PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-324094
(43)Date of publication of application : 24.11.2000

(51)Int.Cl. H04L 9/08
G06F 13/00
G06F 17/30
H04L 9/32

(21)Application number : 2000-025411
(22)Date of filing : 02.02.2000

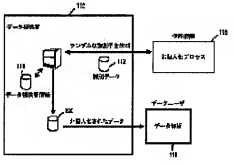
(71)Applicant : SMITHKLINE BEECHAM CORP
(72)Inventor : KOHAN MARK
LANGER DENNIS

(30)Priority

Priority number	Priority date	Priority country
99 118429	02.02.1999	US
99 382127	24.08.1999	US
99 171743	22.12.1999	US

(54) DEVICE AND METHOD FOR MAKING INFORMATION UNINDIVIDUALIZED

(57)Abstract:
PROBLEM TO BE SOLVED: To relate many identification data items to one individual without spoiling the capability of identifying the individual by generating two data sets by separating identification data from other data, and relating an identifier generated by providing only the individual identification information for a trust institution to the other data and generating unindividualized data.
SOLUTION: A data provider 112 processes information inputted in the form of a database 111, separates identification data from data in data provider information 111, and sends the identification data 113 to a trust institution CTTP 116. The trust institution 116 sends back individual data having records including unique identifiers. The data provider 112 matches the unique identifiers to data with the inputted data provider information 111 and separates the unique identifiers related to other information into an unindividualized database 120. The unindividualized database is sent thereafter to a unique data user 118 for analysis.



[0006]

[Means for Solving the Problems] This invention concerns a method to be implemented on a computer and an apparatus, that allow owners or providers of information that incorporate personal identifiers (data providers) to distribute the data to data users in a depersonalized form. "In a depersonalized form" means "without revealing the identity of the person to whom the data relates." The data is otherwise unchanged. Under the method of this invention, the data provider separates personal data from the remainder of the data and creates two data sets. Only personal identifying information is provided to the trusted third party (TTP). The TTP generates identifiers that can be substituted for all the data in the database that can be used to identify individuals, such as names, addresses and social security numbers. The TTP then processes the identifying information by collecting and storing the personal identifying information so that it can later tell whether identifiers generated by the data provider or the TTP relate to the same individual. The data provider relates the identifiers supplied by the TTP with the other data, and generates depersonalized data. The depersonalized data can be sent to data users for analysis. In this way, the data user can match separate records from multiple data providers with a single individual, and the data provider can guarantee that it will not distribute personal identifying information that can link a specific data record with an individual.

[0007]

[Embodiments of the Invention] To put it briefly, this invention is a method and an apparatus for processing confidential information that identifies individuals, allowing anonymous analysis of the data. In the embodiments of this invention explained below, the data provider in possession of a database that contains confidential information divides the information into two parts, identifying information and other information. Using the identifying information, the provider generates a unique identifier for its own use. The unique identifier is linked with the identifying information in the data provider's database. After this, the data owner tags the other information mentioned above with the unique identifier and supplies the tagged data to the data user. In the embodiments set out below, the unique identifier is generated by or registered with the trusted third party (TTP). The trusted third party (TTP) can match the identifying information received from the data provider with other identifying information already in the TTP's database. The TTP is an agency that is under a contractual agreement to protect the identifying information from disclosure and, on the other hand, to maintain and process the data as necessary. By matching the identifying information, the TTP can link multiple identifiers that have been connected to data from several providers. These links can be provided directly to data users, and the data users can correlate data from multiple sources.

[0008] In this invention, the word 'depersonalization' is used to designate the processing step where identifying information is deleted from user data records and is replaced with unique identifiers. This word, as it is used in the technical field of data processing, includes the terms 'anonymization' and

'coding'. When data is anonymized or coded, all identifying information is deleted from the record and a truly random identifier is allocated to refer to the relevant person. In addition, the word 'depersonalization' also includes the processing step of replacing personal identifying information in the data record with an identifier that is not truly random. This type of identifier may be, for example, a hash function value generated from a specified subset of the identifying information, or some other value.

[0009] Figure 1 is a high-level data flow diagram 110 of an exemplary information network that can use the principles of this invention. In this example, a data provider 112 owns or controls a database 114. The database 114 is organized as, for example, several data records. Each record includes at least one data field. Data for each person is stored as a single record, or is linked over several records. A field or part of a field in each record includes data that can be used to identify individuals, i.e. personal identifiable attributes. These attributes include, for example, 'name', 'address' and 'social security number'. Note that these are examples, and are not intended to be a complete list of all identifiable attributes.

[0010] In addition to the identification of information, the database also includes other information about individuals. "Other information" may include, for example, medical information, financial data, buying information and website navigation data. Identifying information may also include non-identifying demographic data, such as a person's occupation, postal code or telephone area code. Depending on the type of other information in the database record, some of this demographic information may be classified as identifying information. For example, if the data records include highly sensitive medical information, the whole postal code may be considered identifying information, but a partial postal code, such as the first three digits of a five-digit postal code, may not be treated as identifying information.

[0011] As the types of information that are considered identifying information vary with the type of data contained in the database, the data provider can decide which pieces of information in the individuals' records are considered identifying information and which pieces of information will be transferred for analysis by the data user. The data provider 112 makes a file 113 from the database. Each record in the file includes fields that have the identifiable attributes from each record in the database. The file 113 is sent to the trusted third party (TTP) 116. The TTP 116 creates unique identifiers linked with the identification attributes. These identifiers may be letters, numbers, a mix of alphanumeric characters, symbols, etc. If the data in the database is highly sensitive, it is possible to generate a unique identifier in a completely random and irreversible fashion, e.g. by taking the instantaneous value of the system clock register. If the data in the database is of low confidentiality, it is possible to generate a unique identifier from the identifying information by a reversible process.

[0012] To generate the unique identifier, the TTP 116 firstly compares the identification data from a record in the file to the records in the internal database 115. The internal database 115 includes

identifying information that has been processed previously by the TTP. Each record in this database also includes a source identifier that identifies the data provider. The data provider owns data relating to the identification record and links it to other records in the database that contain matching identifying information. If the TTP can find a match in its internal database and the previous data source is the provider of the current data, the TTP 116 uses the previously allocated unique identifier as the identifier for the new data. If the source for the previous data is not the provider of the current data, or if the TTP cannot find a match for the data in its database, a new unique identifier will be generated for the data set. Each of the unique identifiers is specific to the data provider.

[0013] By allocating separate unique identifiers to represent the same person with different data providers, the TTP ensures that one data provider cannot identify data owned by another provider. Each data provider has identifying information for all the people within its database, and so if the same unique identifiers were used across multiple providers, one provider could link its identifying information and identify information relating to depersonalized data owned by another data provider. In this way, the confidentiality of the data would be lost.

[0014] When it extracts or generates a unique identifier, the TTP stores the identifier in the appropriate record field in the file 113. Once all of the records have been processed, the TTP116 returns the file 113 to the data provider 112. The data provider generates a new database 120 that includes the records from the original database. Identifiable attributes are deleted from the original database and are replaced with the unique identifiers. The database 120 includes random identifiers that are based on data that have been determined not to have personal identifying attributes, and the database 120 is sent to the data user 118. The data user will have obtained useful data that has been depersonalized, but it will not have the ability to identify individuals that match a particular data set.

[0015] For highly sensitive data, it is desirable that the TTP 116 protects the relationship between the personal identifying information and the unique identifiers. The random identifiers provided by the TTP 116 for this type of information are ideally random as a whole. Apart from the data provider 112 and the TTP 116, no-one can relate the identifiers to particular individuals. The data provider 112 has the authority to grant permission, and only in situations where special permission has been granted will the data user be able to obtain the identifying information concerning the arbitrary information in its possession.

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号
特開2000－324094
（P2000－324094A）
(43)公開日　平成12年11月24日(2000.11.24)

(51)Int.Cl.[7]	識別記号	Ｆ１		テーマコード（参考）
Ｈ０４Ｌ　9/08		Ｈ０４Ｌ	9/00	６０１Ｄ
Ｇ０６Ｆ　13/00	３５１	Ｃ０６Ｆ	13/00	３５１Ｚ
17/30			15/40	３２０Ａ
Ｈ０４Ｌ　9/32		Ｈ０４Ｌ	9/00	６７３Ｃ

審査請求　未請求　請求項の数18　ＯＬ　外国語出願　（全 43 頁）

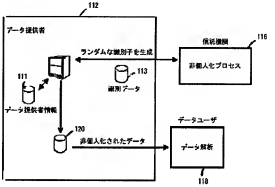
(21)出願番号　特願2000－25411(P2000－25411)
(22)出願日　平成12年２月２日(2000.2.2)
(31)優先権主張番号　６０／１１８４２９
(32)優先日　平成11年２月２日(1999.2.2)
(33)優先権主張国　米国（ＵＳ）
(31)優先権主張番号　０９／３８２１２７
(32)優先日　平成11年８月24日(1999.8.24)
(33)優先権主張国　米国（ＵＳ）
(31)優先権主張番号　６０／１７１７４３
(32)優先日　平成11年12月22日(1999.12.22)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　591002957
スミスクライン・ビーチャム・コーポレイション
SMITHKLINE BEECHAM CORPORATION
アメリカ合衆国ペンシルベニア州19406－0939，キング・オブ・プルシア，スウェードランド・ロード709番
(74)代理人　100062144
弁理士　青山　葆　（外１名）

最終頁に続く

(54)【発明の名称】　情報を非個人化する装置および方法

(57)【要約】　（修正有）
【課題】　個人の識別子を含むデータの所有者または提供者が、データに関連付けられた個人の身元を明らかにすることなく、データを配信できる方法等を提供する。
【解決手段】　データプロバイダは、個人情報を他のデータと分離して、２つのデータセットを生成した後、個人識別情報を信託機関（ＴＴＰ）に提供する。ＴＴＰは一意の識別子を識別情報と関連付け、これでデータベース内のデータを置換する。ＴＴＰはまた、識別情報を処理できるよう個人識別情報を収集し、格納する。ＴＴＰは将来的に識別情報を獲得し、データプロバイダまたはＴＴＰにより生成された一意の識別が同一の個人を示すかどうかを決定する。データプロバイダは自身の一意の識別子またはＴＴＰにより提供された識別子を他のデータに関連付け、非個人化を生成する。非個人化データは解析のためにデータユーザに送られる。



【特許請求の範囲】

【請求項１】　データプロバイダと、データユーザと、信託機関とを含む情報ネットワークにおいて、識別情報フィールドと他のデータフィールドとを含むデータレコードであって、前記レコードの各々の識別情報は１個人を識別するデータレコードの配信方法は、

ａ）前記データレコードの各々について、前記識別情報フィールドと前記他のデータフィールドとを分離して識別レコードを生成するステップと、

ｂ）前記識別レコードのコピーを前記信託機関に伝送するステップと、

ｃ）前記信託機関が、前記識別レコードの各々を一意の識別子に関連付けるステップであって、別個の前記一意の識別子のそれぞれは、１以上の前記識別レコードにより識別される各個人に割り当てられる、ステップと、

ｄ）前記信託機関が、前記一意の識別子を前記データプロバイダに伝送するステップと、

ｅ）前記データプロバイダが、前記他のデータフィールドを前記一意の識別子のそれぞれに関連付けて、非個人化されたデータを形成するステップと、

ｆ）前記データプロバイダの各々が、前記非個人化されたデータを前記データユーザに伝送するステップとからなる、データレコードの配信方法。

【請求項２】　前記信託機関が前記識別レコードの各々を関連付けるステップは、前記一意の識別子として前記識別情報フィールドを元に戻すのに利用できない、ランダムな識別子を生成するステップを含む、請求項１に記載の方法。

【請求項３】　複数のデータプロバイダと、データユーザと、信託機関とを含む情報ネットワークにおいて、識別情報フィールドと他のデータフィールドとを含むデータレコードであって、前記レコードの各々の識別情報は１個人を識別するデータレコードの配信方法は、

ａ）前記複数のデータプロバイダの各々が、前記データレコードの各々について、前記識別情報フィールドと前記他のデータフィールドとを分離して識別レコードを生成するステップと、

ｂ）前記複数のデータプロバイダの各々が、前記識別レコードのコピーを前記信託機関に伝送するステップと、

ｃ）前記信託機関が、前記識別レコードの各々を一意の識別子に関連付けるステップであって、別個の前記一意の識別子のそれぞれは、１以上の前記識別レコードにより識別される各個人に割り当てられる、ステップと、

ｄ）前記信託機関が、前記一意の識別子を前記複数のデータプロバイダのそれぞれに伝送するステップであって、前記複数のデータプロバイダのそれぞれから、一意の識別子を生成するのに用いられた識別レコードが受け取られる、ステップと、

ｅ）前記複数のデータプロバイダの各々が、前記他のデータフィールドを前記一意の識別子のそれぞれに関連付けて、非個人化されたデータを形成するステップと、

ｆ）前記データプロバイダの各々が、前記非個人化されたデータを前記データユーザに伝送するステップとからなる、データレコードの配信方法。

【請求項４】　前記信託機関が、前記識別レコードの各々を一意の識別子に関連付けるステップは、前記一意の識別子として前記識別情報フィールドを元に戻すのに利用できないランダムな識別子を生成するステップであって、前記複数のデータプロバイダの１つ以上により提供される識別情報フィールドは、それぞれ１個人に対応し、別個の一意の識別子は１以上の情報プロバイダのそれぞれについて生成される、請求項１～３のいずれかに記載の方法。

【請求項５】　前記信託機関が、前記識別レコードの各々を一意の識別子に関連付けるステップは、

ａ）前記信託機関が、各個人の相関を記録するステップであって、前記個人に対して複数の一意の識別子が割り当てられ、相関情報を形成するステップと、

ｂ）前記信託機関が、前記データユーザに相関情報を伝送するステップとをさらに含む、請求項１～４のいずれかに記載の方法。

【請求項６】　前記信託機関が、前記データユーザに前記相関情報を伝送するステップは、

ａ）前記データユーザから、前記複数のデータプロバイダのうちの特定のデータプロバイダに対する相関情報の要求を受け取るステップと、

ｂ）前記複数のデータプロバイダのうちの特定のデータプロバイダについての前記相関情報のみを伝送するステップとを含む、請求項１～５のいずれかに記載の方法。

【請求項７】　複数のデータプロバイダと、データユーザと、信託機関とを含む情報ネットワークにおいて、識別情報フィールドと他のデータフィールドとを含む複数のデータレコードであって、前記レコードの各々の識別情報が１個人を識別するデータレコードの配信方法は、

ａ）前記データプロバイダの各々が、前記複数のデータレコードの前記識別情報フィールドから複数の第１の一意の識別子を生成するステップと、

ｂ）前記データプロバイダの各々が、前記複数のデータレコードの各々からの識別情報フィールドのコピーと、前記複数の一意の識別子の各々のコピーとを、複数の識別レコードのそれぞれとして、前記信託機関に伝送するステップと、

ｃ）前記データプロバイダの各々が、前記複数のデータレコードの各々からの他のデータフィールドのコピーと、前記複数の一意の識別子の各々のコピーとを、複数のデータレコードのそれぞれとして、前記データユーザに伝送するステップと、

ｄ）前記信託機関が、前記識別レコードの各々を、第２の一意の識別子と関連付けるステップであって、異なる第２の一意の識別子のそれぞれは、１以上の前記識別レ

コードにより識別される各個人に割り当てられる、ステップと、

ｅ）前記信託機関が、前記第１の一意の識別子と前記第２の一意の識別子とを前記データユーザに伝送するステップと、

ｆ）前記データユーザが、前記データプロバイダにより提供される前記他のデータレコードを前記信託機関により提供される前記一意の識別子に関連付けるステップとからなる、データレコードの配信方法。

【請求項８】　複数のデータレコードを処理し、配信する方法であって、前記複数のデータレコードの各々は、信託機関が１個人を識別するのに用いる情報を含み、

ａ）複数のデータプロバイダから、前記複数の識別レコードのコピーを受け取るステップと、

ｂ）前記識別レコードの各々を、一意の識別子に関連付けるステップであって、別個の一意の識別子は１以上の識別レコードにより識別される各個人に割り当てられる、ステップと、

ｃ）前記複数のデータプロバイダにより提供される前記識別レコードから、特定の個人に関連付けられたレコードを一致させ、前記複数のデータプロバイダにより提供される全ての識別レコードについて同一である第２の一意の識別子を生成する、ステップと、

ｄ）前記一意の識別子を前記データプロバイダのそれぞれに伝送するステップであって、前記データプロバイダのそれぞれから、前記一意の識別子を生成するのに用いられる前記識別レコードを受け取る、ステップとからなる、データレコードの配信方法。

【請求項９】　データプロバイダと、データユーザと、信託機関とを含む汎用コンピュータネットワークであって、前記ネットワークは、識別情報フィールドと他のデータフィールドとを含む複数のデータレコードにアクセスし、前記レコードのそれぞれに含まれる前記識別情報は１個人を識別するネットワークであり、前記ネットワークに

ａ）前記データレコードの各々について、前記識別情報フィールドと前記他のデータフィールドとを分離して識別レコードを生成するステップと、

ｂ）前記識別レコードのコピーを前記信託機関に伝送するステップと、

ｃ）前記信託機関が、前記識別レコードの各々を一意の識別子に関連付けるステップであって、別個の前記一意の識別子のそれぞれは、１以上の前記識別レコードにより識別される各個人に割り当てられる、ステップと、

ｄ）前記信託機関が、前記一意の識別子を前記データプロバイダに伝送するステップと、

ｅ）前記データプロバイダが、前記他のデータフィールドを前記一意の識別子のそれぞれに関連付けて、非個人化されたデータを形成するステップと、

ｆ）前記データプロバイダの各々が、前記非個人化されたデータを前記データユーザに伝送するステップとを行わせる命令セットを含む媒体。

【請求項１０】　前記信託機関が前記識別レコードの各々を関連付けるステップは、前記一意の識別子として前記識別情報フィールドを元に戻すのに利用できない、ランダムな識別子を生成するステップを含む、請求項９に記載の媒体。

【請求項１１】　複数のデータプロバイダと、データユーザと、信託機関とを含む汎用コンピュータネットワークであって、前記ネットワークは、識別情報フィールドと他のデータフィールドとを含む複数のデータレコードにアクセスし、前記レコードのそれぞれに含まれる前記識別情報は１個人を識別するネットワークであり、前記ネットワークに

ａ）前記複数のデータプロバイダの各々が、前記データレコードの各々について、前記識別情報フィールドと前記他のデータフィールドとを分離して識別レコードを生成するステップと、

ｂ）前記複数のデータプロバイダの各々が、前記識別レコードのコピーを前記信託機関に伝送するステップと、

ｃ）前記信託機関が、前記識別レコードの各々を一意の識別子に関連付けるステップであって、別個の前記一意の識別子のそれぞれは、１以上の前記識別レコードにより識別される各個人に割り当てられる、ステップと、

ｄ）前記信託機関が、前記一意の識別子を前記複数のデータプロバイダのそれぞれに伝送するステップであって、前記複数のデータプロバイダのそれぞれから、一意の識別子を生成するのに用いられた識別レコードが受け取られる、ステップと、

ｅ）前記複数のデータプロバイダの各々が、前記他のデータフィールドを前記一意の識別子のそれぞれに関連付けて、非個人化されたデータを形成するステップと、

ｆ）前記データプロバイダの各々が、前記非個人化されたデータを前記データユーザに伝送するステップとを行わせる命令セットを含む媒体。

【請求項１２】　前記信託機関が、前記識別レコードの各々を一意の識別子に関連付けるステップは、前記一意の識別子として前記識別情報フィールドを元に戻すのに利用できないランダムな識別子を生成するステップであって、前記複数のデータプロバイダの１つ以上により提供される識別情報フィールドは、それぞれ１個人に対応し、別個の一意の識別子は１以上の情報プロバイダのそれぞれについて生成される、請求項１１に記載の媒体。

【請求項１３】複数のデータプロバイダと、データユーザと、信託機関とを含む汎用コンピュータネットワークであって、前記ネットワークは、識別情報フィールドと他のデータフィールドとを含む複数のデータレコードにアクセスし、前記データレコードのそれぞれに含まれる前記識別情報は１個人を識別するネットワークであり、前記ネットワークに、

ａ）前記データプロバイダの各々が、前記複数のデータレコードの前記識別情報フィールドから複数の第１の一意の識別子を生成するステップと、

ｂ）前記データプロバイダの各々が、前記複数のデータレコードの各々からの識別情報フィールドのコピーと、前記複数の一意の識別子の各々のコピーとを、複数の識別レコードのそれぞれとして、前記信託機関に伝送するステップと、

ｃ）前記データプロバイダの各々が、前記複数のデータレコードの各々からの他のデータフィールドのコピーと、前記複数の一意の識別子の各々のコピーとを、複数のデータレコードのそれぞれとして、前記データユーザに伝送するステップと、

ｄ）前記信託機関が、前記識別レコードの各々を、第２の一意の識別子と関連付けるステップであって、異なる第２の一意の識別子のそれぞれは、１以上の前記識別レコードにより識別される各個人に割り当てられる、ステップと、

ｅ）前記信託機関が、前記第１の一意の識別子と前記第２の一意の識別子とを前記データユーザに伝送するステップと、

ｆ）前記データユーザが、前記データプロバイダにより提供される前記他のデータレコードを前記信託機関により提供される前記一意の識別子に関連付けるステップとからなる方法を行わせる命令セットを含む媒体。

【請求項１４】　前記複数のデータプロバイダにより提供される前記識別レコードから、特定の個人に関連付けられたレコードを一致させ、前記複数のデータプロバイダにより提供される全ての識別レコードについて同一である第２の一意の識別子を生成する、ステップを行わせる命令をさらに含み、一致させ、生成するステップは、前記信託機関により行われる、請求項１３に記載の媒体。

【請求項１５】　複数のデータレコードの各々が、信託機関によって１個人を識別するのに用いられる情報を含んでおり、前記複数のデータレコードにアクセスする汎用コンピュータに、

ａ）第１のデータプロバイダから、前記複数の識別レコードを受け取るステップと、

ｂ）前記複数の識別レコードの各々を一意の識別子に関連付けるステップであって、別個の一意の識別子のそれぞれは、１以上の前記複数の識別レコードにより識別される各個人に割り当てられる、ステップと、

ｃ）前記一意の識別子を前記データプロバイダに伝送するステップとを行わせる命令セットを含む媒体。

【請求項１６】　前記識別レコードを関連付けるステップは、前記一意の識別子として前記複数の識別情報フィールドを元に戻すのに利用できないランダムな識別子を生成するステップを含む、請求項１５に記載の媒体。

【請求項１７】　複数のデータレコードの各々が、信託機関によって１個人を識別するのに用いられる情報を含んでおり、前記複数のデータレコードにアクセスする汎用コンピュータに、

ａ）複数のデータプロバイダから、前記複数の識別レコードのコピーを受け取るステップと、

ｂ）前記識別レコードの各々を、一意の識別子に関連付けるステップであって、別個の一意の識別子は１以上の識別レコードにより識別される各個人に割り当てられる、ステップと、

ｃ）前記複数のデータプロバイダにより提供される前記識別レコードから、特定の個人に関連付けられたレコードを一致させ、前記複数のデータプロバイダにより提供される全ての識別レコードについて同一である第２の一意の識別子を生成する、ステップと、

ｄ）前記一意の識別子を前記データプロバイダのそれぞれに伝送するステップであって、前記データプロバイダのそれぞれから、前記一意の識別子を生成するのに用いられる前記識別レコードを受け取る、ステップとを行わせる命令セットを含む媒体。

【請求項１８】　前記信託機関が、前記識別レコードの各々を一意の識別子に関連付けるステップは、前記一意の識別子として前記識別情報フィールドを元に戻すのに利用できないランダムな識別子を生成するステップであって、前記複数のデータプロバイダの１つ以上により提供される識別情報フィールドは、それぞれ１個人に対応し、別個の一意の識別子は１以上の情報プロバイダのそれぞれについて生成される、請求項１７に記載の媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、特定の個人と関連付けられたデータの非個人化に関し、特に個人的データを開示することなく、複数のソースからのデータを非個人化する方法に関する。

【０００２】

【従来の技術】現代社会では、特定の個人に関する情報が数多くの団体から得られる。病院、研究所、銀行、保険会社および小売業者などの健康、金融および営利団体は、研究および開発、マーケティング、および他の商業目的のために利用可能なデータを所有している。しかし、このようなデータに関係する個人のプライバシー保護が必要であるとの意識が高まってきている。特に、個人の健康または財政状態に関する情報は、きわめて機密性が高いといえる。

【０００３】この情報を分析するには、多くの場合、複数のソースからデータにアクセスする必要がある。例えば、特定の薬物療法の効能を判定する研究には、薬物療法を処方する介護人のグループや、薬物を処方する薬局の該当するグループのレコードにアクセスする必要がある。

【０００４】

【発明が解決しようとする課題】各データプロバイダが所有するデータには機密性の高い情報が含まれており、それらのデータは、その情報を解析できるデータユーザと共有できないようすべきである。多くのデータプロバイダは、自らのデータから任意の識別情報を削除して医学データのみをデータユーザに提供できる。それに対し、データユーザはさまざまなソースからのデータを相関させることができないので、分析に必要とされるであろう情報をも失うことになる。

【0005】したがって、個人データに関連付けられたその個人の識別能力を損うことなく、しかし多数のソースからの個人データ項目を1個人に関係するとして関連付ける能力を備えた、さまざまなソースからの個人データを獲得する方法が必要になってきている。

【0006】

【課題を解決するための手段】本発明は、コンピュータに実装される方法と、装置とに関し、個人の識別子を含むデータの所有者または提供者（データプロバイダ）が、データユーザに非個人化された形式でそのデータを配信できるようにする。「非個人化された形式で」とは、すなわちそのデータに関連付けられた個人の身元を明らかにすることなく、ということである。その他の点ではそのデータは変更されない。本発明の方法によれば、データプロバイダは他のデータから個人データを分離して、2つのデータセットを作成する。個人識別情報のみが、信託機関（TTP（Trusted Third Party））に提供される。TTPは、名前、住所または社会保障番号といった個人を識別するのに利用可能なデータベース内の全データを置換する識別子を生成する。TTPはまた、個人識別情報を収集、格納し、それによりTTPは識別情報を処理し、データプロバイダまたはTTPによって生成された識別子が同一の個人に関連するか否かの判断を将来得ることができる。データプロバイダはTTPにより提供された識別子を他のデータに関連付け、非個人化データを作成する。非個人化データは、分析のためにデータユーザに送ることができる。このようにして、データユーザは、1個人に関連する1以上のデータプロバイダからの別個のレコードを一致させることができ、データプロバイダは、ある個人と特定のデータレコードとをリンクする個人識別情報を配信しないことを保証できる。

【0007】

【発明の実施の形態】本発明は、端的にいえば、個人を識別する機密情報を処理する方法および装置であり、これにより匿名でのデータ解析に利用可能となる。以下に説明される本発明の実施の形態では、機密情報を含むデータベースを所有するデータプロバイダは、その情報を2つの部分、すなわち識別情報と他の情報とに分割する。識別情報を用いて、プロバイダは、自身のために固有識別子を生成する。固有識別子は、データプロバイダのデータベース内の識別情報にリンクされる。その後、データ所有者はこの固有識別子を上述した他の情報にタグ付けし、データユーザはこのタグ付けされたデータを提供する。以下に説明する各実施の形態では、固有識別子は信託機関（TTP）により生成され、または信託機関に登録されている。信託機関（TTP）は、データプロバイダから受け取った識別情報と、すでにTTPのデータベースに存在している別の識別情報とを一致させることができる。TTPは、識別情報が開示されることから保護し、その一方で必要に応じてそのデータを保守および処理するという、契約による同意の下にある機関である。識別情報を一致させることにより、TTPは、多数のプロバイダからのデータに関連付けられた複数の識別子をリンクできる。これらのリンクは直接データユーザに提供され、データユーザは複数のソースからのデータを相関させることができる。

【0008】本発明では、「非個人化」という語は、識別情報がユーザデータレコードから削除され、固有識別子により置換される処理を説明するのに用いられる。この語は、データ処理の技術分野で用いられるように、「匿名化」や「符号化」という語をも含む。データが匿名化または符号化されると、すべての識別情報がレコードから削除され、真にランダムな識別子がその人物を表すように割り当てられる。加えて、「非個人化」という語はまた、真にランダムでない識別子とデータレコード内の個人識別情報とを置換する処理をも含む。このタプの識別子は、例えば、識別情報の所定の部分集合から生成されたハッシュ関数値または他の値である。

【0009】図1は、本発明の原理を利用できる例示的な情報ネットワークのハイレベルデータフロー図110である。この例においては、データプロバイダ112はデータベース114を所有または制御する。データベース114は、例えば、複数のデータレコードとして組織されている。各レコードは、1以上のデータフィールドを含む。各人のデータは単一レコードとして保持され、または複数のレコードにわたってリンクされている。各レコードのフィールドまたはフィールドの部分は、個人を識別するのに利用可能なデータ、すなわち個人識別可能属性を含む。これらの属性は、例えば、「名前」、「住所」および「社会保障番号」を含む。なお、これは例示であり、識別可能な属性を網羅して列挙したものではない。

【0010】情報の識別に加えて、データベースは個人についての他の情報も含む。この「他の情報」は、例えば、医療情報、財政データ、購買情報またはウェブサイトナビゲーションデータを含むことができる。識別情報はまた、非識別性人口統計データ、例えば、ある人の職業、郵便番号または電話のエリアコードを含むことができる。データベースレコード内の「他の情報」のタイプによって、この人口統計情報のいくつかは識別情報とし

て分類できる。例えば、データレコードが機密性の高い医療情報を含む場合には、全郵便番号が識別情報として考慮され、部分的な郵便番号、例えば、5桁の郵便番号の上位3桁は識別情報とはならない。

【0011】識別情報であると考えられる情報のタイプはデータベース内に格納されたデータのタイプとともに変化するので、データプロバイダは、個人のレコードにあるどの情報が識別情報であると考えられているのか、およびデータユーザの解析のためにどの情報が渡されるのかを決定できる。データプロバイダ112は、データベースからファイル113を作成する。ファイルの各レコードは、データベース内の各レコードからの識別可能属性を有するフィールドを含む。ファイル113は、信託機関（TTP）116に送られる。TTP116は、識別属性と関連付けられた固有識別子を作成する。この識別子はアルファベット、数値、英数字、記号等である。データベース内のデータの機密性が高い場合には、全くランダムに、そして、例えばシステムクロックレジスタの瞬時値を取ることによって不可逆的に固有の識別子を生成できる。データベース内のデータの秘密性が低い場合には、可逆的なプロセスによって識別情報から固有の識別子を生成できる。

【0012】固有の識別子を生成するために、TTP116はまずファイル内のレコードから内部データベース115内のレコードまでの識別データを比較する。内部データベース115は、あらかじめTTPにより処理された識別情報を含む。このデータベースの各レコードはまた、データプロバイダを識別するソース識別子を含む。データプロバイダは識別レコードに関連するデータを所有し、一致する識別情報を含むデータベースの中の別のレコードにリンクする。TTPがその内部データベースで一致を見出さない場合、かつ前のデータのソースが現在のデータの供給者である場合には、TTP116は、前に割り当てられた一意の識別子を新たなデータの識別子として用いる。前のデータのソースが現在のデータの供給者でないか、またはTTPがそのデータベース内のデータに一致を見出さない場合には、新たな一意の識別子がそのデータセットに対して生成される。一意の識別子の各々は、そのデータプロバイダに固有のものである。

【0013】別個の一意の識別子を割り当てて、別個のデータプロバイダのそれぞれにおいて同一人物を表すことにより、TTPは、あるデータプロバイダが別のプロバイダによって所有されるデータを識別できないことを確実にする。各データプロバイダはそのデータベース内のすべての人々について識別情報を有するので、仮に複数のプロバイダにおいて同一の一意の識別子が用いられている場合には、あるプロバイダは識別情報をリンクして、別のデータ供給者が所有する非個人化されたデータに関する情報を識別できる。これによりデータの秘密性がなくなることになる。

【0014】一意の識別子を取り出しまたは生成すると、TTPはその識別子をファイル113内の適当なレコードフィールドに格納する。すべてのレコードが処理されると、TTP116はファイル113をデータプロバイダ112に戻す。データプロバイダは、もとのデータベースのレコードを含む新たなデータベース120を生成する。もとのデータベースからは識別可能な属性は除去され、一意の識別子に置換される。データベース120は、個人識別属性でないと判断されたデータに基づくランダムな識別子を含み、データベース120はデータユーザ118に送られる。データユーザは非個人化された有用なデータを得ることになるが、特定のデータセットに一致する個人を識別する能力は有さない。

【0015】機密性の高いデータに対して、TTP116は個人識別情報と一意の識別子との間の関係を保護することが望ましい。このタイプの情報については、TTP116により提供されるランダムな識別子は望ましくは全体としてランダムである。データプロバイダ112またはTTP116以外は、誰も識別子をある個人に関連させることはできない。データプロバイダ112が許可する権限を持ち、特別の許可を出した状況においてのみ、データユーザはその所有する任意のデータについて識別情報を得ることができる。

【0016】本発明の例示的な実施の形態においては、個人は、データプロバイダによって所有され、制御されているデータベース内の複数のレコードを有する。加えて上で述べたように、TTP116は、複数のデータプロバイダからのある人物に関するデータを有している。新たに受け取られた個人データをすでにデータベース115に存在するデータにリンクするために、TTP116は、受け取ったデータに一致アルゴリズムを実行する。データyが複数のプロバイダからのデータを要求する場合には、TTP116が必要となる。

【0017】本発明では、多くの一致アルゴリズムが利用可能である。例示的な一致アルゴリズムは、M. A. Jaroによる、"Probabilistic Linkage of Large Public Health Data Files"（Statistics in Medicine, vol. 14, John Wiley, pp 491-498 (1995)）と題された論文に開示され、また、I. P. Fellegi らによる "A Theoryof Record Linkage"（Journal of the American Statistical Association, vol. 64, No. 328, pp 1183-1210 (1969)）と題された記事に開示されている。最も簡単な一致アルゴリズムは、決定一致法である。このアルゴリズムによれば、新たに受け取られた個人データからの個人データフィールドは、データベース115からのデータに存在する、対応するフィールドと比較される。これらのすべてのフィールドが一致すると、新たに受け取られたデータは、ほぼ確実にデータベース内にデータが存在する個人のデータである。決定一致法に用いることができる例示的なフィールドセットは、ラストネーム、ファ

ーストネーム、アドレスおよび社会保障番号である。電話番号や誕生日といった別のフィールドも利用できる。

【0018】しかし決定一致技術は、不完全なデータまたは書き換えエラーに起因して、2つのデータベース間のすべての一致や、高いパーセンテージの一致を識別できない。決定一致技術を拡張する1つの方法は、確率技術を利用して2つの類似しないフィールドが一致する尤度を判断することである。別の技術は、例えば、決定一致を行う前、または確率技術を適用する前に、省略個所またはニックネームを拡充することによりデータを正規化することである。さらに別の方法は、編集経過により一致しないレコード内の類似しないフィールドを分析して、書き換えの際に生じ得るエラーを識別することである。

【0019】ある1つのデータ一致技術を以下に説明する。この方法は、1999年11月15日に提出された、現在係属中の米国出願第60/165,121号に開示されており、使用可能な多くの考えられる一致法の1つである。この出願に開示されている内容は、本発明の理解および実施に関する内容である限りにおいて、ここに引用することにより組み込まれる。例示的な一致技術は、3ステップからなり、i) データの標準化、ii) 重みの評価、iii) データの比較である。

【0020】定義

以下の定義および略記が、例示的な本実施の形態について利用される。

【0021】μ=確率：式（1）により与えられるように、任意のランダムな要素対が偶然に一致する確率。

$$\mu = \frac{n_{match}}{n_{A*B}} \qquad (1)$$

【0022】ρ=確率：データ要素の信頼性。要素エラー率（EER：Element Error Rate）が0．99以上の場合は、ρ＝1－EERであり、それ以外の場合は、ρ＝0．99－EERとなる。

【0023】一致：所与の要素対が正確に合致し、両方の要素が

$$A_{e_1} = B_{e_1}$$

であると知られている状態。

【0024】一致重み：式（2）に示すような、レコード一致プロセス中に要素対が合致した場合に要素対に割り当てられる重み。

$$\mathrm{AW} = -\log_2\left(\frac{\rho}{\mu}\right) \qquad (2)$$

【0025】デカルト積：順序対の集合A＊B＝{（a，b）|a∈A∧b∈B}。

【0026】不一致：所与の要素対が正確には合致しておらず、両方の要素が

$$A_{e_1} \neq B_{e_1}$$

として知られている状態。

【0027】不一致重み：式（3）に示すような、レコード一致プロセス中に要素対が合致しない場合に要素対に割り当てられる重み。

$$\mathrm{DW} = \log_2\left(\frac{1-\rho}{1-\mu}\right) \qquad (3)$$

【0028】要素エラー率：少なくとも1つの要素が未知、例えばヌルである、式（4）に示すような要素対の比率。

$$\varepsilon = \frac{n_{null}}{n_{A*B}} \qquad (4)$$

【0029】頻度テーブル：回数の一覧であり、現れる変数の異なる値すべてのパーセンテージ。

【0030】平均：式（5）に示すような、算術平均。

$$\bar{x}_i = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (5)$$

【0031】非決定：所与の要素対における要素のいずれか1つまたは両方が未知である状態。

【0032】ランダム数割り当て：本発明の例示的な実施の形態において、約1500のvブロックがR＝int（（U＊P）＋1）により生成されるようデータセット内のあらゆるレコードにはランダム数が割り当てられる。ここでRは結果として得られるランダム数を、Uは（以下に定義する）上界を、Pは0と1の間の値を返すランダム関数である。本発明の例示的な実施の形態において、Pは擬似ランダム整数発生器とすることができる。

【0033】閾値：確率一致法で利用される閾値は、$-\infty \leq x \leq \infty$ の範囲をもつ、2進数字確率比である。

【0034】上界：式（6）に示されるような、データセットを1500のほぼ等しい行に分割する層の数。

$$U = \mathrm{int}\left(\frac{NR}{1500}\right) \qquad \mathrm{NR}：データセット内のレコード数 \quad (6)$$

【0035】このプロセスで用いられるコンピュータおよびマシン言語に関して、短いオーダで相当に多い量の計算を実行できるハードウェアであれば、どのようなものでも要求を満たす。現在の水準の任意のPCまたはサーバが利用できる。オペレーティングシステムに関しては、UNIX（登録商標）が好ましいが、Windows（登録商標）に関するWindows98やNT等も利用できる。ソースコードは任意の言語で記述されてよく、好ましい場合にはJava（登録商標）でもよい。

【0036】データの標準化

このプロセスの第1ステップは、入力されたファイルのデータの標準化を含む。この標準化は、更なる精密さおよび信頼性のために必要とされる。入力されたファイルは、任意の数の変数を含む。1以上の変数は、例えば、個人などの特定のデータソースに対して一意であり、または一意であってよい。有用な変数の例は、メンバ識別

子、運転免許証番号、社会保障番号、保険会社コード番号、氏名、性別、誕生日、通りの住所、町、州、郵便番号、市民権である。加えて、いくつかの識別子はさらに、その基本的または根本的な要素をさらに引き出すことができる。例えば、氏名はファーストネーム、ラストネームおよびミドルイニシャルの根本的な要素に分けられる。

【００３７】標準化プロセスの間、全ての文字データは好ましくは単一の格に変換され、全ての略語またはニックネームは、より長い形式に変換される。例えば、全ての文字は大文字に変換される。したがって、例えば、ファーストネームは大文字に標準化され、｛ＢＯＢ，ＲＯＢ，ＲＯＢＢＹ｝＝ＲＯＢＥＲＴとなる。町および通りの一般名称は郵便番号に変換され、例えば、米国では米国郵政公社規格に変換される。後者の例では、この変換は、工業規格ＣＡＳＳが認証したソフトウェアを用いて行われる。

【００３８】重み評価

$$e_n=\begin{cases}1 & A_{e_n}=B_{e_n}\text{のとき} \quad (\text{一致})\\ 0 & A_{e_n}=\text{Null および／または}B_{e_n}=\text{Null} \quad (\text{決定なし})\\ -1 & A_{e_n}\neq B_{e_n}\text{のとき} \quad (\text{不一致})\end{cases} \qquad (7)$$

ここで、A_{e_n}はレコードＡからのｎ番目の要素である。

【００４１】この行列が一旦完全にアクセスされると、各e_nのパーセンテージは作表され、格納される。このプロセスは反復数（例えば、１５）の値だけ繰り返される。

【００４２】同意および決定なしの平均パーセンテージが、各データ要素について計算される。各データ要素のρ一確率、または信頼性は、式（８）のように計算される。

$$\varepsilon=\overline{x}_{\text{Percent No Decisions}} \quad (\text{決定なしの確率の平均})\ \text{とすると}$$

$$\rho=\begin{cases}\text{if } \varepsilon \geq .99 \text{ then } 1-\varepsilon\\ \text{else } .99-\varepsilon\end{cases} \qquad (8)$$

【００４３】μ一確率、または任意の所与のレコード対についての要素ｎが偶然一致する確率は、式（９）のように計算される。

$$\mu=\overline{x}_{\text{Percent Agreement}} \quad (\text{一致の確率}) \qquad (9)$$

【００４４】ρ一確率およびμ一確率から、式（１０）および（１１）の各々を用いて、不一致重み公式および一致重み公式が計算できる。

$$\text{不一致}=\log_2\left(\frac{1-\rho}{1-\mu}\right) \qquad (10)$$

$$\text{一致}=\log_2\left(\frac{\rho}{\mu}\right) \qquad (11)$$

【００４５】一意の識別子の割り当て本プロセスの最終段階は、入力されたデータセットの中での、一意に識別するもの（entities）の動作（action）である。

この例示的なアルゴリズムの根本となる要素は、確率関数に必要な一致重みおよび不一致重みを評価するプロセスである。重みは、対話式ブートストラップ技術を用いて、見込みが一致する（chance agreement）確率に基づいて計算される。

【００３９】例示的な重み評価プロセスの第１ステップは、データセットを１５００行のほぼ等しいブロックに分割するのに必要な層の数を決定することである（図２の２０１－２１９、式（６）参照）。

$$U=\text{int}\left(\frac{NR}{1500}\right) \qquad \text{NR：データセット内のレコード数} \qquad (6)$$

【００４０】その後、ソースファイルはスキャンされ、レコードには１からＵの間のランダム数が割り当てられる。そしてデータ行列が生成される。データ行列は、割り当てられたランダム数を持ったレコードのデカルト積を含んでいる。そうしてえられた行列がスキャンされる。各レコード対の中にある各要素はアクセスされ、式（７）に示される値が割り当てられる。

【００４６】入力されたファイルからのレコードの各々は、参照データベース１１５について評価され、それによりデータにより表されるもの（entity）が決定一致技

術および確率一致技術を組み合わせて用いて前に識別されているかを判断する。データにより表されるものが参照セットの中にすでに表されていると判断されると、入力されたレコードには、一意した参照レコードから一意の識別子（ＵＩＤ：unique identifier）が割り当てられる。データにより表されるものがまだ参照セットには存在しない場合には、新たなＵＩＤがランダムに生成され、割り当てられる。ランダムな値は多くの別個のアルゴリズムを用いて生成できる。上で説明したように、データの機密性が高い場合には、ランダムな識別子が真にランダムであることが望ましい。ランダムな識別子は、例えば、システムクロックレジスタの即値を用いて生成される。それほど機密性が高くない場合には、可逆的な方法を用いてもよい。しかし識別子は一意であることが好ましく、任意のある１つの識別子には１人だけが関連付けられるべきである。このランダムな識別子は、数値、英数字または記号（例えば、空間的なパターンまたはホログラム）であってよい。

【００４７】ＵＩＤ割り当てが起こると、入力されたレコードはその全体において評価され、それによりそのレコードが参照テーブルにいまだ含まれないものを一意に表現するかが判断される。新たなレコードである場合には、将来的な使用のために、そのレコードは参照データベース１１５に挿入される。

【００４８】決定一致技術

例示的な決定一致技術は単にブール論理を利用するものであり、データが標準化された後に適用される。２つのレコードが、ある基準を満たしているかに適合するかが判断される。例えば、以下のようになる。

【００４９】ファーストネームが正確に一致しているか、ラストネームが正確に一致しているか、誕生日が正確に一致しているか、社会保障番号またはメンバー識別子が正確に一致しているか、である。

【００５０】２つのレコードが決定一致法の基準を満足する場合には、確率的処理は行わない。しかし、決定一致が生じない場合には、入力されたレコードは、確率一致法に付される。

【００５１】確率一致技術

確率一致プロセスの第１ステップは、入力されたレコードの固有の要素の特徴に基づいて、参照テーブルから広報レコードセットを構築することである。このプロセスはブロッキングと呼ばれ、候補レコードはブロッキングテーブルと呼ばれる。全てのデータセットに同一の文字を用いることはなく、本プロセスで用いられる要素はデータ解析を経て決定される。しかしながらブロッキング変数は、個人に関してやや一意な要素を含むのがよい。やや一意とは、例えば、社会保障番号、または誕生日およびラストネームの組み合わせである。

【００５２】ブロッキングテーブルの構築が完成すると、各候補レコードの各要素は、入力されたレコードからの対応する要素と比較される。計算手順についての式（１２）を参照されたい。

$$W_n = \begin{cases} \text{一致重み} & A_{e_n} = B_{e_n} \text{のとき} \\ 0 & A_{e_n} = \text{Null および/または} B_{e_n} = \text{Null} \\ \text{不一致重み} & A_{e_n} \neq B_{e_n} \text{のとき} \end{cases} \qquad (12)$$

ここで、A_{e_n}はレコードＡからのｎ番目の要素である。

【００５３】そして式（１３）に示すように、全ての候補レコードについて複合重みが計算される。

$$W = \sum_{i=1}^{N} W_i \qquad (13)$$

【００５４】その後、最も高い複合重みがかかった候補レコードがあらかじめ定められた閾値に対して評価される。重みが閾値に一致するまたは超えると、候補レコードは入力されたレコードに適合すると判断される。重みが閾値を超えない場合は、入力されたレコードは、いまだ参照セットに含まれていないものを表すと想定される。

【００５５】例示的な一致技術は、不一致の２つのフィールドが同一のデータを表すかを決定しようとする技術ではない。例えば、書き換えエラーのために社会保障番号１２３４５６７８９が１２３４５６７９８と記録されているとすると、上で説明したアルゴリズムは不一致を示す。上で説明したアルゴリズムに対する別の拡張では、類似のフィールド間の編集距離（ＥＤ：Ｅｄｉｔ Ｄｉｓｔａｎｃｅ）ような、類似性の基準を利用できる。例えば、上述の社会保障番号は、編集距離１である。その理由は、最後の２桁の桁を置換すると正確な結果になるからである。この類似性の基準は、例えば、確率プロセスの一部として、または確率プロセスの結果が正しいことを確認するための後の処理ステップとして利用できる。

【００５６】図２、３、４および５は、データプロバイダ１１２からデータユーザ１１８へ機密性の高い情報を匿名で伝送する際に、ＴＴＰ１１６を利用する別の実施

の形態を示す。各実子の形態は、単一のデータプロバイダを含むが、図２を除いて、全ての実施の形態が複数の独立した情報プロバイダを含むよう拡張可能であることが意図されている。図２に示す実施の形態は、単一の情報プロバイダからの複数の情報ソースを含んでもよい。ある実施例は、図６を参照して以下に説明するように複数の情報プロバイダが示されている。

【００５７】図２に示す実施の形態では、データ提供者１１２は、データベース１１１で入力された情報を処理し、データベース内のデータから個人データ１１３を分離する。個人データは上で説明した処理のために、ＴＴＰ１１６に送られる。ＴＴＰ１１６は、一意の識別子を含むこととなった各レコードを備えた個人データを返す。そしてデータ提供者１１２は、入力されたデータベース１１１で一意の識別子をデータに一致させ、他の情報と関連付けられた一意の識別子を非個人化されたデータベース１２０に分離する。この非個人化されたデータベースは、その後解析のため一意データユーザに送られる。

【００５８】図２に示す例示的な実施の形態では、ＴＴＰ１１６とデータユーザ１１８との間には、直接的な通信はない。この実施の形態は、単一のデータプロバイダが複数のデータソースを含み、様々なデータソースからのデータを一致させル必要がある場合に利用できる。この例の１つは、料金請求レコード、患者治療法レコード、薬局レコード、放射線医学レコードおよび治療レコードが別々に、おそらく別々の請負業者に保持されている病院環境である。病院は、自身の使用のために内部的にこれらのレコードを一致させたり、外部のデータユーザにデータを提供することを強く望むであろう。本実施の形態では、ＴＴＰ１１６は様々なデータソースからのレコードを一致させ、すべてソースの間で各個人についての単一の一意の識別子を提供する。

【００５９】図３に示す例示的な本実施の形態は、以下の点で図２に示すものと異なる。すなわち、ＴＴＰ１１６はデータプロバイダに一意の識別子を通信しないことである。本実施の形態では、プロバイダ１１２は入力されたデータベースを処理して、２つのデータベースを生成する。データベースの１つ１１３は、識別情報のみを有し、他方のデータベースはその他の情報のみを有する。データプロバイダは共通の識別子を２つのデータベースに存在する対応するレコードに割り当てる。これらの識別子はレコード番号のような単純なものや、特定の個人に対するランダムな識別子のような複雑なものであってよい。第１の例では、データプロバイダは同一人について複数のレコードをリンクするよう試みることはない。第２の例では、データプロバイダはレコードをすでにリンクしており、データプロバイダはデータベース１１３のレコードとデータベース１２０の対応するレコードの双方に、その個人について一意の識別子を設ける。データプロバイダが割り当てた一意の識別子は、ランダム、擬似ランダムまたは可逆であってもよい。しかし、可逆的な一意の識別子は、少なくともいくつかの個人情報が開示される状況においてのみ利用できる。

【００６０】データベース１１３はＴＴＰ１１６に提供される。データベース１１６は、同一の識別情報を有するレコードが互いに一致するよう、および同一の識別情報を有するレコードがＴＴＰ１１６の（図示されない）内部データベースのレコードに一致するよう、上述のように処理される。

【００６１】同時に、識別データがＴＴＰに送られ、別のデータを含むデータベース１２０がデータユーザ１１８に送られる。データベース１２０を受け取ると、データユーザはＴＴＰ１１６から相関するデータ３１０を受け取るのを待つ。この相関するデータは、データプロバイダからのレコード識別子群または一意の識別子群を、ＴＴＰにより生成された一意の識別子群に一致させる。データユーザはＴＴＰ１１６により生成された一意の識別子群をデータベース１２０の適当なレコードに加え、ＴＴＰに一意の識別子を用いてその他の情報を処理する。

【００６２】図３に示すシステムが複数のデータプロバイダに用いられる場合には、ＴＴＰ１１６により提供される相関するデータ３１０はまた、複数のデータプロバイダにより提供される一意の識別子またはレコード番号の間の関係を示すテーブルを含む。この情報を用いて、データユーザ１１８はデータ解析の前に、複数のプロバイダからのデータを関連付けることができる。図４に示すシステムは、図２を参照して上に説明したシステムと以下の点で異なり、その他の点では類似である。すなわち異なるのは、図４のシステムでは、ＴＴＰ１１６とデータユーザ１１８との間に通信があることである。図４では、データ提供者は識別情報をＴＴＰ１１６に送る。ＴＴＰ１１６はデータを一致させ、一意の識別子を加え、その一意の識別子を有する識別情報をデータ提供者に送り返す。そしてデータ提供者は、識別情報レコードから関連付けられた別の情報レコードにいたるまでの一意の識別子をコピーし、データユーザにその他の情報レコードを提供する。データユーザ１１８はそれから相関するデータ４１０をＴＴＰ１１６から直接受け取る。この例では、相関する情報は他のデータ提供者からの一意の識別子を含む。この一意の識別子は、データ提供者１１２により提供される非個人化データ１２０の一意の識別子と対応する。

【００６３】図４に示すシステムでは、相関するデータ４１０は、データプロバイダ１１２の要求があるとＴＴＰ１１６によりデータユーザ１１８に提供され、またはデータユーザ１１８により要求される。データプロバイダからデータが要求されると、ＴＴＰはそのデータベース内のデータ提供者の全てに対し、相関する情報を提供

する。しかしデータユーザがデータを求めると、データユーザは、データを受け取るデータプロバイダのみからの情報を要求することになる。

【0064】図5のシステムは、図3を参照して上に説明したシステムと以下の点で異なり、その他の点では類似である。すなわち異なるのは、TTP116がデータユーザに全ての相関するデータを送るというよりは、TTP116は特定の要求にのみ応答してデータユーザに相関するデータを送ることである。図4に示すシステムに関しては、その要求は、データユーザ118にデータを提供するデータプロバイダのみついての要求である。

【0065】図1～5に示すシステムのいずれにおいても、データユーザは、データを評価している個人の識別が必要である。例えば、データユーザ118が医療データを処理し、生命をおびやかす条件を割り出したとすると、データユーザはその個人に告知することが必要となる。この場合、データユーザはデータ提供者にデータ識別情報を求めることができる。データユーザに用いられている一意の識別子がデータプロバイダにより保持されている識別子に一致しない場合には、データプロバイダ112はTTP116に権限を与え、データユーザ118に情報を明かさせる。

【0066】図6は、本発明の原理を用いる別の例示的な実施の形態を示す。本実施の形態では、信託機関116は各データプロバイダ112a、112bおよび112cに、非個人化プロセスを行うソフトウェアおよび／またはハードウェアと、識別された非個人化されたデータを保持する補助データベース115a、115bおよび115cを提供する。補助データベース115a、115bおよび115cの各々は、データプロバイダ112a、112bおよび112cに対する個人の識別可能な属性および個人の識別子を含む。個人の識別可能な属性および個人の識別子は、TTP116により所有され、制御される中央データベース115から得ることができる。中央データベース115には、後の処理の間に、権限を与えられたのような情報のソースから得られる情報が存在する。データプロバイダは、各レコードをデータユーザ118に提供することを希望し、データプロバイダはレコードのｎ記フィールドを抽出し、それらを非個人化プロセスに入力する。非個人化プロセスは、データユーザにより保持されている情報を、信託機関（TTP）により提供されたデータベースに以前に格納されている情報と一致させることにより、ランダムな識別子を割り当てる。一致データがそれぞれのデータベース115a、115bおよび115cに見つからない場合には、一意であり、かつランダムであり得る識別子が割り当てられ、そのプロセスからの出力として提供される。前に非個人化されたデータとの一致が起こると、最初に割り当てられた一意の識別子はそのプロセスからの出力として提供される。データプロバイダ112a、112bおよび112cは、一意の識別子をレコード内の個人を識別可能属性に取り換え、個々の非個人化されたレコードを生成する。そしてデータ提供者は非個人化されたレコードをデータユーザ118に送る。

【0067】非個人化されたデータの複数のソースへのリンクを可能にするため、各データプロバイダ112a、112bおよび112cは、データプロバイダの非個人化プロセス116a、116bおよび116cにより割り当てられた識別データおよび一意の識別子を含むファイルをTTP116に提供する。TTPはこれらのファイルを相関させることにより各データプロバイダにより提供される識別情報レコード間の一致を割り出し、何らかの相関がある表示とともに一意の識別子を中央データベースに格納する。データプロバイダにより権限が与えられると、TTPは他のデータプロバイダからのランダムな識別子を示すデータユーザに、情報を提供できる。ランダムな識別子は同一の個人に関連し、したがってデータユーザは、リンクされた非個人化されたデータベース120を生成できる。

【0068】ある例では、データプロバイダ112aは識別データをTTP116に提供することはない。しかし本例では、TTP116は、電話帳のような公衆ソースからのデータが前もって設けてある中央データベースを保守し、データプロバイダに一致アルゴリズムを提供する。そしてTTP116は、以前にTTP116のデータベースと一致していたデータ提供者からのファイルのみを受け取る。子供のような、公衆のデータベースには存在しない、ある個人のグループ内でのデータの相関は、データユーザから除外できる。しかし、このプロセスは偽陽性（false positive）相関の方が好ましい。偽陰性（false positive）相関よりも偽陽性（false positive）相関の方が好ましい。

【0069】当業界の実務者であれば、本発明の基本的概念の多くの変更が認識できるであろう。すなわち、データプロバイダおよびデータユーザを有する信託機関の利用により、データがプロバイダからユーザへ移動するときにデータを非個人化できる。上述した実施の形態は、本来例示的なものであり、本発明が実施される様々なやり方を網羅して列挙したものではない。

【0070】図7は、図1～6に示す任意の情報ネットワークの、例示的な物理的実施形態のブロック図である。例示的なシステムは、ローカルエリアネットワークまたはワイドエリアネットワーク716によりリンクされている。これらのネットワークはまた、ダイレクト通信インターフェース718およびリムーバブルメディア722によって、インターネットのようなグローバル情報ネットワークにも接続されている。図7の例示的なシステムは、6つの処理システム710、730、740、760、770および780を含む。これらのシステム710、730、740、760、770および7

８０の各々は、関連付けられたデータベース７１２、７３２、７４２、７６２、７７２および７８２を有する。データプロバイダ、データユーザおよびＴＴＰにより保守されるデータベースは、当業界において現在周知の、市販されている入手可能なホストコンピュータ上に存在する。

【００７１】例示的な処理システム７１０は、ホストコンピュータ７１４と、ネットワークインターフェース７１６とを含む。ホストコンピュータ７１４はネットワークインターフェース７１６によりローカルエリアネットワーク、ワイドエリアネットワークまたはグローバル情報ネットワークを介して、他のデータ処理システムと通信できる。図７に示すように、ホストコンピュータ７１４は、ローカルエリアネットワーク（ＬＡＮ）７１７を介して処理システム７４０、７３０と通信する。コンピュータ７１４はまた、ＬＡＮ７１７を用いてグローバル情報ネットワークサーバ７５０と通信し、さらにサーバ７５０およびグローバル情報ネットワーク７５２を経てリモートユーザ７６０、７８０と通信する。ネットワークインターフェースに加え、処理システム７１０のホストコンピュータ７１４は、通信インターフェース７１８、例えばモデムを含む。モデムを経て、処理システム７１０はリモートユーザ７７０と通信する。処理システム７１０はまた、入力／出力（Ｉ／Ｏ）プロセッサ７２０を含む。入力／出力（Ｉ／Ｏ）プロセッサ７２０は、リムーバブルメディア装置７２２が結合されている。リムーバブルメディア装置７２２は、例えばディスケットドライブであり、これを経ることにより、ホストコンピュータは、ホストコンピュータ７１４と直接的または間接的なデータ通信経路を持たない他のコンピュータシステムと通信できる。

【００７２】各ホストコンピュータは、１以上のプロセッサ（図示せず）と、メモリ（図示せず）と、入力および出力装置（図示せず）と、大容量記憶媒体（図示せず）へのアクセス手段を含む。各処理システムは、現在当業界において周知の、単一システムまたはコンピュータネットワークであってもよい。データプロバイダ、ＴＴＰおよびデータユーザは、ＬＡＮ７１７のようなコンピュータネットワークで、またはある位置から別の位置までリムーバブルメディア７２２にデータを物理的に転送することにより、データを交換できる。システムはまた、インターネットなどのグローバル情報ネットワークにわたって実現可能である。ホストコンピュータおよびグローバル情報ネットワークはまた、複数のリモートユーザと通信できる。

【００７３】「データベース」という語は、レコードおよびフィールド、またはそれらと同等のものを用いた任意のデータベースを意味するものとして、広く解釈される。その手法は、データのコード化に用いられるハイレベル言語により限定されることはなく、または要求されたデータ処理を実現するプログラムのコード化に用いられる言語によっても限定されない。本発明は、データプロバイダ１１２、信託機関１１６およびデータユーザ１１８によって実行されるコンピュータソフトウェアで実施されることが意図されている。このコンピュータソフトウェアは、ディスケット、ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭ等の媒体上、または、無線周波数または音声周波数の搬送波上に実現される。

【００７４】図８および９は、本発明の例示的な実施の形態を示すフローチャートである。図８は、図６において行われる処理を示し、図９は、図３、４または５において行われる処理を示す。

【００７５】図８において、ステップ８１０では、ＴＴＰ１１６は２つの小売店（小売店１１２ａおよび小売店１１２ｂ）に符号化プロセスおよび符号化データベースを提供する。小売店は、その店の中で処理およびデータベースを実施する。本発明の例示的な実施の形態において、ＴＴＰ１１６により提供されるデータベース１１５ａおよび１１５ｂには、ＴＴＰ中央データベース１１５から提供された情報が前もって存在している。提供された情報には一意の識別子は含まれない。

【００７６】ステップ８１２において、小売店１１２ａおよび１１２ｂの各々は、個人の人口統計属性および個人の識別子を、各データレコードから抽出する。これらは小売店１１２ａおよび１１２ｂの各々がデータユーザ１１８（本例では出荷取次店）に送りたいと考えるデータである。各レコードについては、情報は、ＴＴＰの提供した符号化プロセスを経て処理される。ステップ８１４において、符号化プロセスは各レコードに一意の識別子を割り当てる。ステップ８１４では次に、小売店１１２ａおよび１１２ｂが個人の人口統計属性および個人の識別子を単一の一意の識別子に置換することにより、非個人化されたデータを生成する。単一の一意の識別子は、符号化プロセスにより提供される。小売店１１２ａおよび１１２ｂは、出荷取次店１１８に非個人化されたデータを送る。続いてステップ８１８において、小売店１１２ａおよび１１２ｂは、各レコードについて割り当てられた一意の識別子をＴＴＰ１１６に送る。各レコードは、符号化プロセスを実行している間、一致が生じる。ステップ８２０において、ＴＴＰ１１６は、小売店１１２ａおよび１１２ｂにより提供された一意の識別子割り当て情報をその中央データベース１１５に格納する。同じくステップ８２０では、ＴＴＰ１１６は、同一の個人にリンクする小売店１１２ａおよび１１２ｂに対する一意の識別子を相関情報として出荷取次店１１８に送る。

【００７７】ステップ８２２では、出荷取次店は相関情報を用いてデータをリンクし、その市場調査を行う。この調査は個人の身元を明らかにする能力を必要とせずに行われる。ブロック８２２から８１２への矢印により示

されるように、このプロセスは繰り返し行われる。周期的に、ＴＴＰ１１６は符号化プロセスに最新情報を送り、小売店１１２ａおよび１１２ｂにはデータベースを送る。これらの更新情報は、ＴＴＰ１１６により得られる符号化プロセスを行う中央データベースへの更新情報／追加情報に起因する。これらの更新情報を処理した後は、小売店１１２ａおよび１１２ｂは全ての一意の識別子をＴＴＰ１１６に送り返す。これらの識別子は、新たに提供された情報に対して小売店により前もって割り当てられたものである。

【００７８】本発明の本実施の形態では、小売店１１２ａおよび１１２ｂは個人を識別可能などのような小売情報も提供することはないことに留意されたい。小売店によって出荷取次店に提供される小売情報には、個人を識別可能な属性は存在しない。したがって、出荷取次店１１８は現実の個人が誰であるかを知ることができない。しかしこのような事情であっても、出荷取次店１１８は、小売店の情報力を利用して市場調査能力を増強できる。図９に示す本発明の例示的な実施の形態では、製造業者１１８は、３つの地方ヘルスケアプロバイダのヘルスケア情報を利用して、特定の病状の健康状態の性質を同定することを欲している。３つのデータプロバイダ１１２（プロバイダＡ、プロバイダＢおよびプロバイダＣ）は個人が誰かを識別する情報（例えば、会員番号、社会保障番号、氏名等）を有する。製造業者１１８、プロバイダＡ、プロバイダＢおよびプロバイダＣは契約により信託機関に権限を与え、図９に示すヘルスケアデータ符号化プロセスを用いてヘルスケアデータを符号化する。

【００７９】このプロセスのステップ９１０では、プロバイダＡ、プロバイダＢおよびプロバイダＣは各々個人を識別可能な情報を、内部データベース１１１から抽出し、ファイル１１３に入れる。ステップ９１２では、プロバイダＡ、プロバイダＢおよびプロバイダＣはファイルをＴＴＰ１１６に送る。

【００８０】ステップ９１４では、ＴＴＰ１１６は一致プロセスを用いて各個人を識別し、各レコードに符号化鍵を割り当てる。ステップ９１６では、ＴＴＰ１１６は対応する符号化鍵を有するファイルを、プロバイダＡ、プロバイダＢおよびプロバイダＣに送り返す。続いてステップ９１６では、プロバイダＡ、プロバイダＢおよびプロバイダＣは、製造業者１１８に送信を希望するレコードの各々に対する個人属性を、ＴＴＰ１１６から受け取られた符号化鍵に置き換える。またステップ９１８では、プロバイダＡ、プロバイダＢおよびプロバイダＣは製造業者１１８に符号化されたヘルスケア情報を送る。ステップ９２０では、製造業者は符号化されたヘルスケア情報ファイルを受け取り、ＴＴＰ１１６から相関するデータを得る。最後にステップ９２２では、製造業者はプロバイダＡ、プロバイダＢおよびプロバイダＣからのデータをリンクし、その調査を終了する。この調査は製造業者がどの個人を同定できるかに関係することなく終了する。

【００８１】以上、多くの例示的な実施の形態によって本発明を説明した。上述したように、本発明は、特許請求の範囲に記載された発明の範囲内で様々に変更して実現できることが企図されている。

【図面の簡単な説明】

【図１】　本発明において、データが様々な機関の間をどのように伝送されるかを説明するのに有用なデータフロー図である。

【図２】　第１の例示的なデータ非個人化方法を示すデータフロー図である。

【図３】　第２の例示的なデータ非個人化方法を示すデータフロー図である。

【図４】　第３の例示的なデータ非個人化方法を示すデータフロー図である。

【図５】　第４の例示的なデータ非個人化方法を示すデータフロー図である。

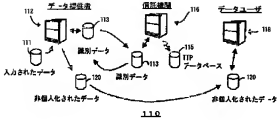
【図６】　複数のデータプロバイダがどのように信託機関と相互作用し、１以上のデータユーザにより相関されたデータを提供するかを示すデータフロー図である。

【図７】　図１～６に記載された方法を実施するために利用可能な、例示的なコンピュータの構成を示すブロック図である。

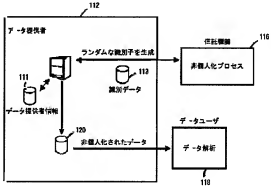
【図８】　図６の例示的な方法のフローチャートである。

【図９】　図３、４または５の例示的な方法のフローチャートである。

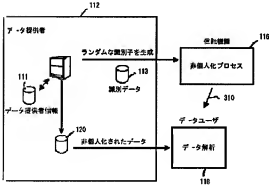
【図1】



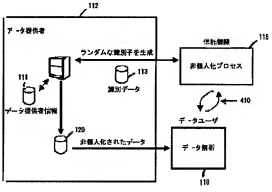
【図2】



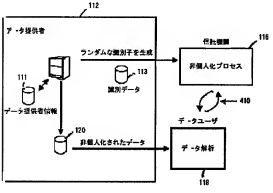
【図3】



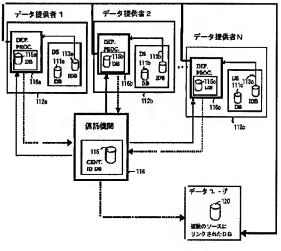
【図4】



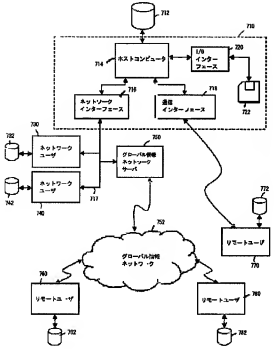
【図5】



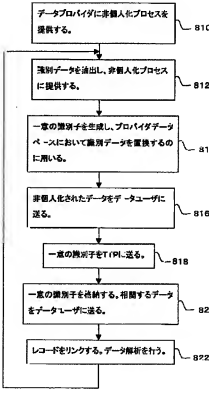
【図6】



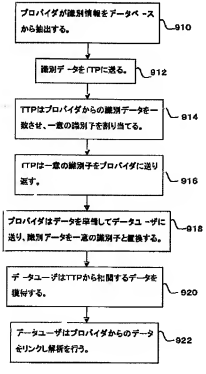
【図７】


712
710
714
ホストコンピュータ
I/O インターフェース
720
716
718
ネットワークインターフェース
通信インターフェース
722
700
732
ネットワークユーザ
730
グローバル情報ネットワークサーバ
734
ネットワークユーザ
740
772
リモートユーザ
770
750
グローバル情報ネットワーク
760
リモートユーザ
762

【図８】


データプロバイダに非個人化プロセスを提供する。
810
識別データを抽出し、非個人化プロセスに提供する。
812
一意の識別子を生成し、プロバイダデータベースにおいて識別データを置換するのに用いる。
814
非個人化されたデータをデータユーザに送る。
816
一意の識別子をTTPに送る。
818
一意の識別子を格納する。相関するデータをデータユーザに送る。
820
レコードをリンクする。データ解析を行う。
822

【図９】


プロバイダが識別情報をデータベースから抽出する。
910
識別データをTTPに送る。
912
TTPはプロバイダからの識別データを一致させ、一意の識別子を割り当てる。
914
TTPは一意の識別子をプロバイダに送り返す。
916
プロバイダはデータを平準してデータユーザに送り、識別データを一意の識別子と置換する。
918
データユーザはTTPから相関するデータを復号する。
920
データユーザはプロバイダからのデータをリンクし解析を行う。
922

フロントページの続き

(72)発明者 マーク・コーハン
アメリカ合衆国19335ペンシルベニア州ダウニングタウン、クリークス・ベンド204番

(72)発明者 デニス・ランガー
アメリカ合衆国08540ニュージャージー州プリンストン、クリーブランド・レイン12番

【外国語明細書】

APPARATUS AND METHOD FOR DEPERSONALIZING INFORMATION

BACKGROUND OF THE INVENTION

The present invention concerns the depersonalization of data associated with a particular individual and, in particular, a method for depersonalizing data from several sources without disclosing the personalized data.

In modern society, information relating to specific individuals is obtained by numerous organizations. Healthcare, financial and commercial organizations such as hospitals, laboratories, banks, insurance companies and retailers own data that could be used for research and development, marketing, and other business functions. There is, however a growing awareness for the necessity to maintain the privacy of the individuals connecting with the data. In particular, information regarding an individual's health or financial status may be extremely sensitive.

The analysis of this information often requires accessing data from multiple sources. For example, a study to determine the effectiveness of a particular medication may need to access records from a group of caregivers that prescribe the medication and from a corresponding group of pharmacies who prescribe the medication. The data owned by each of the data providers contains sensitive information that they may be unable to share with the data user who will be analyzing the information. While the various data providers could remove any identifying information from their data and provide only the medical data to the data user, the data user would not be able to correlate the data from the various sources and, thus, would lose information that would be needed in the analysis.

Therefore, a need has arisen for a method for obtaining personal data from multiple sources without the ability to identify the individual associated with the data but with the ability to associate individual data items from multiple sources as relating to a single individual.

SUMMARY OF THE INVENTION

The present invention relates to a computer implemented method and apparatus that allows an owner or provider of data that contains personal identifiers (data provider) to distribute that data to a data user in a depersonalized form, i.e., without revealing the identity of the individuals associated with the data. The data is otherwise unchanged. According to this method, a data provider separates the personal information from the other data to create two data sets. Only the personal identifying information is provided to a Trusted Third Party (TTP). The TTP generates an identifier that replaces any data in the database that can be used to identify an individual, such as name, address or social security number. The TTP may also collect and store the personal identifying information so that it can process identifying information that it acquires in the future to determine if the identifiers generated by the data provider or by the TTP refer to the same individual. The data provider associates the identifier provided by the TTP with the other data to create depersonalized data that may be sent to a data user for

analysis. In this manner, different records from one or more data providers that refer to a single individual can be matched by the data user, and the data provider is assured that no personal identifying information is distributed that would link an individual to a particular data record.

DETAIL DESCRIPTION OF THE DRAWINGS

Figure 1 is a data flow diagram which is useful for describing how data is transferred among the various parties in the subject invention.

Figure 2 is a data flow diagram which illustrates one exemplary data depersonalization method.

Figure 3 is a data flow diagram that illustrates a second exemplary data depersonalization method.

Figure 4 is a data flow diagram that illustrates a third exemplary data depersonalization method.

Figure 5 is a data flow diagram that illustrates a fourth exemplary data depersonalization method.

Figure 6 is a data flow diagram that shows how multiple data providers may interact with a trusted third party to provide data that may be correlated by one or more data users.

Figure 7 is a block diagram that shows an exemplary computer configuration that may be used to implement the methods described in Figures 1 through 6.

Figure 8 is a flow-chart diagram of an exemplary method of Figure 6.

Figure 9 is a flow-chart diagram of an exemplary method of Figures 3, 4 or 5.

DETAILED DESCRIPTION OF THE INVENTION

Briefly, the present invention is a method and apparatus for processing sensitive information, that identifies a person, so that it may be used for anonymous data analysis. In the embodiments of the invention described below, a data provider, who owns a database containing sensitive information, divides the information into two parts, identifying information and other information. Using the identifying information, the provider generates, or has generated for it, a unique identifier that is linked to the identification information in the data provider's database. The data owner then tags the other information with this unique identifier and provides the tagged data to the data user. In each of the embodiments described below, the unique identifier is generated by or registered with a Trusted Third Party (TTP) who is able to match the identifying information received from the data provider to other identifying information that may already be in the TTP's database. A TTP is an entity that is under a contractual agreement to protect the identifying information from being disclosed, while maintaining and processing the data as necessary. By matching the identifying information, the TTP can link identifiers that are associated with data from multiple providers. These links may be provided directly to the data users to allow the data users to correlate data from multiple sources.

In the subject application, the word "depersonalizing" is used to describe the process by which the identifying information is removed from a user data record and replaced by a unique identifier. This term encompasses the terms "anonymizing" and "encoding" as they are used in the data processing arts. When data is anonymized, or encoded, all identifying information is removed from a record and a truly random identifier is assigned to represent the person. In addition, the term "depersonalizing" also encompasses a process by which an identifier that is not truly random is replaces the personal identifying information in a data record. An identifier of this type may be, for example, a hash function value or other value produced from a predetermined subset of the identifying information.

Fig. 1 shows a high-level data flow diagram of an exemplary information network, 110, with which the principles of the present invention may be used. In this exemplary embodiment, a data provider 112 owns or controls a database, 114, which, for example, is organized as a plurality of data records, each record containing one or more data fields. The data for each person may be kept in a single record or it may be linked across multiple records. Fields or portions of the fields in each record contain data that can be used to identify the individual, namely, personal identifiable attributes. These attributes include, for example, "name," "address" and "social security number". This is an exemplary and not exhaustive listing of the identifiable attributes.

In addition to the identifying information, the database contains other information about the individual. This "other information" may include, for example, medical information, financial data, purchase activity information or web-site navigation data. The identifying information may also include non-identifying demographic data, for example, the person's occupation, their postal code or their telephone area code. Depending on the type of "other information" in the database record, some of this demographic information may be classified as identifying information. For example, if the data record includes sensitive medical information then the entire postal code may be considered identifying information while a partial postal code, for example the first three digits of a five-digit zip code, would not be identifying information.

Because the type of information that may be considered to be identifying information varies with the type of data stored in the database, the data provider is best able to decide which information in the person's record is considered to be identifying information and which information may be passed on to a data user for analysis. The data provider 112 creates a file 113 from the database, each record of the file contains the fields having the identifiable attributes from each record in the database. The file 113 is sent to a Trusted Third Party (TTP) 116. The TTP 116 creates a unique identifier to be associated with the identifying attributes. This identifier can be alphabetic, numeric, alphanumeric, symbolic and the like. If the data in the database is sensitive, the unique identifier may be generated in a totally random fashion and in a manner that cannot be reversed, for example by taking the instantaneous value

of the system clock register. If the data in the database is less confidential, the unique identifier may be generated from the identifying information by a reversible process.

To generate the unique identifier, the TTP 116 first compares the identifying data from a record in the file to records in an internal database 115 that contains identifying information which has previously been processed by the TTP. Each record of this database also contains a source identifier that identifies the data provider, who owns the data associated with the identifying record, and links to other records in the database that contain matching identifying information. If the TTP finds a match in its internal database and if the source of the previous data is the supplier of the current data then the TTP 116 uses the previously assigned unique identifier as the identifier for the new data. If the source of the previous data was not the supplier of the current data or if the TTP does not find a match for the data in its database a new unique identifier is generated for the data set. Each unique identifier is specific to the data provider.

By assigning a different unique identifier to represent the same person for respectively different data providers, the TTP ensures that one data provider can not identify any data owned by another provider. Because each data provider has identifying information for all of the people in its database, if the same unique identifier were used for multiple providers, one provider could link its identifying information to depersonalized data that is owned by a different data supplier. This may result in a breach of confidentiality for that data.

After retrieving or creating the unique identifier, the TTP stores it into a field of the appropriate record in the file113. When all of the records have been processed, the TTP 116 returns the file 113 to the data provider 112. The data provider creates a new database 120 containing the records of the original database from which the identifiable attributes are removed and replaced with the unique identifier. The database 120 containing the random identifiers along with the data not determined to be personal identifying attributes are then sent to the data user 118. The data user now has useful data that has been depersonalized so that the data user does not have the ability to identify an individual that matches a particular set of data.

For sensitive data, it is desirable for the TTP 116 to protect the relationship between the personal identifying information and the unique identifiers. For this type of information, the random identifiers provided by the TTP 116 are desirably totally random; there should be no way for anyone other than the data provider 112 or the TTP 116 to relate the identifier with the individual. Only in the circumstance where the data provider 112 has authority to grant and grants specific permission should the data user be able to obtain identifying information for any data in its possession.

In this exemplary embodiment, an individual may have multiple records within the database owned or controlled by the data provider. In addition, as set forth above, the TTP 116 may have data on one

person from multiple data providers. In order to link newly received personal data to data already in the database 115, the TTP 116 executes a matching algorithm on the data that it receives. In any scenario in which a data user requires data from multiple providers, a TTP 116 is necessary.

Many matching algorithms may be used in the present invention. Exemplary matching algorithms are disclosed in a paper by M A. Jaro entitled "Probabilistic Linkage of Large Public Health Data Files" Statistics in Medicine, vol. 14, John Wiley, pp 491-498 (1995) and in an article by I. P. Fellegi et al. entitled "A Theory of Record Linkage" Journal of the American Statistical Association, vol. 64, No. 328, pp 1183-1210 (1969). The simplest matching algorithm is a deterministic match. By this algorithm, individual data fields from the newly received personal data are compared to corresponding fields in the data from the database 115. If all of these fields match, then the newly received data is almost certainly for the person whose data is in the database. An exemplary set of fields that may be used for a deterministic match are Last Name, First Name, Address and Social Security Number. Other fields such as Telephone Number and Birth Date may also be used.

Deterministic matching techniques may not identify all matches or even a large percentage of matches between two databases because of incomplete data or transcription errors. One method for enhancing deterministic matching techniques is to employ probabilistic techniques to determine the likelihood that two dissimilar fields match. Another technique is to normalize the data, for example by expanding abbreviations and nicknames before performing the deterministic match or applying the probabilistic techniques. Yet another method is to analyze dissimilar fields in otherwise matching records by their edit distances to identify possible errors in transcription.

One exemplary data matching technique is presented below. This method is disclosed in copending U.S. patent application No. 60/165,121 filed 15 November 1999 and is one of many possible matching methods that may be used. The materials disclosed therein are incorporated by reference herein to the extent they are material to the understanding and practice of this invention. The exemplary matching technique comprises three steps, i) data standardization, ii) weight estimation, and iii) data comparison.

Definitions

The following definitions and abbreviations are used for this exemplary embodiment:

μ-Probability: The probability that any random element pair will match by chance, as given by equation (1).

$$\mu = \frac{n_{match}}{n_{obs}} \qquad (1)$$

ρ-Probability: The reliability of the data element. If the Element Error Rate is ≥ .99 then $\rho = 1 - EER$; Else $\rho = .99 - EER$

Agreement: A condition such that a given element pair matches exactly and both elements are known $A_{e_1} = B_{e_1}$

Agreement Weight: The weight assigned to an element pair when they agree during the record matching process as shown in equation (2).

$$AW = -\log_2\left(\frac{\rho}{\mu}\right) \quad (2)$$

Cartesian Product: The set of ordered pairs $A * B = \{(a,b) \mid a \subset A \wedge b \subset B\}$

Disagreement: A condition such that a given element pair does not exactly match and both elements are known $A_{e_1} \neq B_{e_1}$

Disagreement Weight: The weight assigned to an element pair when they disagree during the record matching process as shown in equation (3).

$$DW = \log_2\left(\frac{1-\rho}{1-\mu}\right) \quad (3)$$

Element Error Rate: The proportion of element pairs where at least one element is unknown, e.g., null, as shown in equation (4).

$$\varepsilon = \frac{n_{null}}{n_{total}} \quad (4)$$

Frequency Table: Summary of the number of times, and percentage of total different values of a variable occur

Mean: Arithmetic average, as given in equation (5).

$$\overline{x_i} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (5)$$

No Decision: A condition such that a given element pair where either one or both of the elements is unknown.

Random Number Assignment: In the exemplary embodiment of the invention, every record in the data set is assigned a random number such that v blocks of approximately 1500 are created $R = \text{int}\{(U * P) + 1\}$ where R is the resulting Random Number, U is the Upper Bound (defined below) and P is a random function that returns a value between 0 and 1. In the exemplary embodiment of the invention, P may be a pseudo random number generator.

Threshold: The threshold utilized in probabilistic matching is a binit odds ratio with a range of $-\infty \geq x \leq \infty$.

Upper Bound: Number of strata such that the data set is divided into approximately equal rows of 1500 as shown in equation (6).

$$v = \text{int}\left(\frac{\text{Number of Records in Data Set}}{1500}\right) \qquad (6)$$

As regards the computer and machine language used in this process, just about any piece of hardware capable of executing a fairly large number of calculations in short order will fill the bill. Any current state-of-the-art PC or server could be used. As for the operating system, UNIX is preferred, but Windows 98 or NT for Windows or the like could be used. The source code can be written in any language, though Java if preferred.

Data Standardization

The first step of this process involves the standardization of data in an input file. This standardization is required for increased precision and reliability. The input file can contain any number of variables of which one or more are or may be unique to a particular data source such as an individual. Examples of useful variables are: member identifier, drivers' license number, social security number, insurance company code number, name, gender, date of birth, street address, city, state, postal code, citizenship. In addition, some identifiers can be further distilled down into their basic, or atomic, components. For example, a name may be broken down into atomic components of first name, last name and middle initial.

During the standardization process, all character data is preferably transformed to a single case, and all abbreviations or nick-names are transformed to their longer forms. For example all letters may be transformed to uppercase. So for instance, first names are standardized to uppercase, e.g., {BOB, ROB, ROBBY} → ROBERT. Common names for cities and streets may be transformed to the postal code, e.g., in the U.S. to United States Postal Service standard. In the latter instance this can be performed using industry standard CASS certified software.

Weight Estimation

A fundamental component of this exemplary algorithm is the process of estimating the agreement and disagreement weights necessary for the probabilistic function. Weights are calculated based in probabilities of chance agreement using an iterative bootstrap technique.

The first step in the exemplary weight estimation process is to determine the number of strata required such that the data set can be divided into approximately equal blocks of 1500 rows (Fig. 2 - 201-219), see equation (6).

$$v = \text{int}\left(\frac{\text{Number of Records in Data Set}}{1500}\right) \qquad (6)$$

The source file is then scanned and the records are assigned a random number between 1 and U. A data matrix is created containing a Cartesian product of records with a random number of 1 assigned.

The resulting matrix is then scanned. Each element pair within each record pair is assessed and assigned a value as shown in equation (7).

$$e_n = \begin{cases} 1 \text{ if } A_{e_n} = B_{e_n} \text{ (Agreement)} \\ 0 \text{ if } A_{e_n} = Null \text{ and/or } B_{e_n} = Null \text{ (No decision)} \\ -1 \text{ if } A_{e_n} \neq B_{e_n} \text{ (Disagreement)} \end{cases} \quad (7)$$

where A_{e_n} is the nth element from record A

Once the matrix has been fully assessed, percentages for each e_n are tabulated and stored. This process may be repeated for a number (e.g. 15) of iterations.

Mean percentages of Agreements and No Decisions are calculated for each data element. The ρ probability, or the reliability, for each data element is then calculated, see equation (8).

$$let\ \varepsilon = \overline{x}_{\text{Percent NoDecision}}$$

$$\rho = \begin{cases} if\ \varepsilon \geq .99\ then\ 1 - \varepsilon \\ else\ .99 - \varepsilon \end{cases} \quad (8)$$

The μ probability, or the probability that element n for any given record pair will match by chance, is calculated see equation (9).

$$\mu = \overline{x}_{\text{Percent Agreement}} \quad (9)$$

From the ρ and μ probabilities, the disagreement and agreement weight formula may calculated employing equations (10) and (11) respectively.

$$Disagreement = \log_2\left(\frac{1-\rho}{1-\mu}\right) \quad (10)$$

$$Agreement = \log_2\left(\frac{\rho}{\mu}\right) \quad (11)$$

Unique Identifier Assignment

The final stage of this process is the action of uniquely identifying entities within the input data set.

Each record from the input file is evaluated against the reference database 115 to determine if the entity represented by the data has been previously identified using a combination of deterministic and probabilistic matching techniques. If it is judged that the entity is already represented in the reference set, the input record is assigned the unique identifier (UID) from the reference record that it has matched against. If it is judged that the entity represented by data is not yet in the reference set, a new UID is randomly generated and assigned. Random values may be generated using many different

algorithms. As set forth above, if the data is sensitive, it is desirable that the random identifier be truly random, generated, for example, using the instantaneous value of the system clock register. For less sensitive data reversible methods may be used. It is desirable, however, for the identifier to be unique; only one person should be associated with any one identifier. This random identifier may be numeric, alphanumeric, or symbolic (e.g. a spatial pattern or hologram).

After the UID assignment occurs, the input record is evaluated, in its entirety, to determine if the record is a unique representation of the entity not already contained in the reference table. If it is a new record, then it is inserted into the reference database 115 for future use.

Deterministic Matching Technique

The exemplary deterministic matching technique employs simple Boolean logic and is applied after the data has been standardized. Two records are judged to match if certain criteria are met, such as the following:

First Name Matches Exactly

Last Name Matches Exactly

Date of Birth Matches Exactly

Social Security Number OR Member Identifier Matches Exactly

If two records satisfy the criteria for deterministic matching, no probabilistic processing occurs. However, if no deterministic match occurs, the input record is presented for a probabilistic match.

Probabilistic Matching Technique

The first step in the probabilistic matching process is to build a set of candidate records from the reference table based on characteristics of specific elements of the input record. This process is referred to as blocking, the set of candidate records is referred to as the blocking table. All data sets do not use the same characteristics, the elements used in this process are determined through data analysis. It is suggested, however, that the blocking variables include those elements that are somewhat unique to an individual, e.g., social security number, or a combination of date of birth and last name.

Upon completion of the construction of the blocking table, each element for each candidate record is compared against its corresponding element from the input record. See equation (12) for the scoring mechanism.

$$w_n = \begin{cases} \textit{Agreement Weight if } A_{e_n} = B_{e_n} \\ 0 \textit{ if } A_{e_n} = \textit{Null and/or } B_{e_n} = \textit{Null} \\ \textit{Disagreement Weight if } A_{e_n} \neq B_{e_n} \end{cases} \tag{12}$$

where A_{e_n} is the nth element from record A

A composite weight is then calculated for all candidate records, see equation (13).

$$W = \sum_{i=1}^{N} W_i \qquad (13)$$

The candidate record with the highest composite weight is then evaluated against a predefined threshold. If the weight meets or exceeds the threshold, the candidate record is judged to match the input record. If the weight does not exceed the threshold, it is assumed that the input record represents an entity not yet included in the reference set.

The exemplary matching technique does not attempt to determine whether two fields that disagree represent the same data. If, for example, because of a transcription error, a social security number of 123 45 6789 were recorded as 123 45 6798, the algorithm set forth above would indicate disagreement. One alternative enhancement to the algorithm set forth above may be to employ some measure of similarity such as Edit Distance between similar fields. For example, the social security numbers described above have an edit distance of one because a digit substitution of the last two digits would produce the correct result. This measure of similarity may be employed, for example, as a part of the probabilistic process or as a post processing step to confirm that the result of the probabilistic process is correct.

Figures 2, 3, 4 and 5 show alternative embodiments for employing a TTP 116 in the anonymous transfer of sensitive information from a data provider 112 to a data user 118. Although each of the embodiments includes a single data provider, it is contemplated that, except for Figure 2, all embodiments may be expanded to include multiple independent information providers. The embodiment shown in Figure 2 may include multiple information sources from a single information provider. One implementation that illustrates multiple information providers is described below with reference to Figure 6.

In the embodiment shown in Figure 2, a data supplier 112 processes input information in the database111 to separate the personal data 113 from the other data in the database. The personal data is sent to the TTP 116 for processing, as described above. The TTP 116 returns the personal data with each record now including a unique identifier. The data supplier 112 then matches the unique identifier to the data in the input database 111 and separates the other information and the associated unique identifiers into a depersonalized database 120. This depersonalized database is then sent to the data user 118 for analysis.

In the exemplary embodiment shown in Figure 2, there is no direct communication between the TTP 116 and the data user 118. This embodiment may be used where a single data provider includes multiple data sources and needs to match the data from the various data sources. One example of this is a hospital environment in which billing records, patient treatment records, pharmacy records, radiology

records and therapy records may be kept separately, perhaps by separate contractors. The hospital may want to match these records internally for its own use and may want to provide the data to an external data user. In this embodiment, the TTP 116 matches the records from the various data sources and provides a single unique identifier for each person among all of the sources.

The exemplary embodiment shown in Figure 3 differs from that shown in Figure 2 in that the TTP 116 does not communicate the unique identifier to the data provider. In this embodiment, the provider 112 processes its input database to generate two databases. One database, 113 has only identifying information and the other database has only the other information. The data provider assigns common identifiers to corresponding records in the two databases. These identifiers may be as simple as a record number or as complex as a random identifier for a particular individual. In the first instance, the data provider makes no attempt to link multiple records for the same person. In the second instance, the data provider has already linked the records and has placed the unique identifier for the person into both the records of the database 113 and the corresponding records of the database 120. Where the data provider has assigned unique identifiers, the identifiers may be random, pseudo random or reversible. It is noted, however, that reversible unique identifiers may only be used in situations where at least some personal information may be disclosed.

The database 113 is provided to the TTP 116 where it is processed, as described above, to match records having the same identifying information to each other and to records in the internal database (not shown) of the TTP 116.

At the same time that the identifying data is sent to the TTP, the database 120 containing the other data is sent to the data user 118. After receiving the database 120, the data user waits to receive correlating data 310 from the TTP 116. This correlating data matches the record identifiers or unique identifiers from the data provider to unique identifiers generated by the TTP. The data user adds the unique identifiers generated by the TTP 116 to the appropriate records of the database 120 and processes the other information using the TTP unique identifiers.

When the system shown in Figure 3 is used with multiple data providers, the correlating data 310 provided by the TTP 116 may also include a table indicating correspondence among the unique identifiers or record numbers provided by the multiple data providers. Using this information, the data user 118 may associate data from the multiple providers before performing the data analysis

The system shown in Figure 4 is similar to that described above with reference to Figure 2 except that, in the system of Figure 4, there is communication between the ITP 116 and the data user 118. In Figure 4, the data supplier sends the identifying information to the TTP 116 who matches the data, adds unique identifiers and sends the identifying information with the unique identifiers back to the data supplier 112. The data supplier then copies the unique identifiers from the identifying information

records to the associated other information records and provides the other information records to the data user 118. The data user 118 then receives correlating data (410) directly from the TTP 116. In this instance, the correlating information includes unique identifiers from other data suppliers that correspond to the unique identifiers in the depersonalized data 120 that is provided by the data supplier 112.

In the system shown in Figure 4, this correlating data 410 may be provided by the TTP 116 to the data user 118 at the request of the data provider 112 or it may be requested by the data user 118. When the data is requested by the data provider, the TTP provides correlating information for all of the data suppliers in its database. When the data user asks for data, however, it requests information from only those data providers from which it receives data.

Figure 5 shows a system that is similar to the system shown in Figure 3 except that, rather than send all correlating data to the data user, the TTP 116 sends correlating data to the data user 118 only in response to a specific request. As with the system shown in Figure 4, that request may be for only those data providers who supply data to the data user 118.

In any of the systems shown in Figures 1 through 5, it may be necessary for the data user to identify the person whose data is being evaluated. If, for example, the data user 118 is processing medical data and identifies a life-threatening condition, the data user may need to notify the individual. In this instance, the data user may ask the data supplier for the identifying information. In situations where the unique identifiers being used by the data user do not match the identifiers held by the data provider, the data provider 112 may then authorize the TTP 116 to divulge the information to the data user 118.

Figure 6 illustrates another exemplary embodiment using the principles of the present invention. In this embodiment, The Trusted Third Party 116 provides each data provider 112a, 112b and 112c with software and/or hardware that performs the depersonalizing process and a supporting database 115a, 115b and 115c that holds the identified depersonalized data. Each database 115a, 115b and 115c contains individual identifiable attributes and individual identifiers for the respective data provider 112a, 112b and 112c obtained from a central database 115 owned or controlled by the TTP 116. The central database 115 is populated with information obtained from authorized sources of such information during past processing. For each record the data provider wishes to supply to a data user 118, the data provider extracts the identifying fields for the record and inputs them into the depersonalizing process. The depersonalizing process assigns the random identifier by matching the information held by the data user with information previously stored in the database provided by the Trusted Third Party. If no matching data is found in the respective database 115a, 115b and 115c, a unique and possibly random identifier is assigned and provided as output from the process. If a match with previously

depersonalized data is encountered, the unique identifier assigned initially is provided as output from the process. The data providers 112a, 112b and 112c substitute the unique identifiers for the individual identifiable attributes in the record to create respective depersonalized records. The data suppliers then send the depersonalized records to the data user 118.

In order to enable the linking of multiple sources of depersonalized data, each data provider 112a, 112b and 112c supplies, to the TTP 116, a file containing the identifying data and the unique identifiers assigned by the data provider's depersonalizing process 116a, 116b and 116c. The TTP correlates these files to identify matches among the identifying information records provided by the respective data providers and stores the unique identifiers, with indications of any correlation, within the central database. When authorized by the data provider, the TTP may supply information to the data user showing the random identifiers from any of the data provider that relates to the same individual, thus allowing the data user to create a linked depersonalized database 120.

In some instances, a data provider 112a will not supply the identifying data to the TTP 116. In this instance, the TTP 116 will maintain a central database that is pre-populated with data from public sources, such as telephone directories, and will supply the matching algorithms to the data provider. The TTP 116 will receive only those files from a data supplier that have been previously matched with the TTP 116 database. It is apparent that correlation of data within certain groups of individuals who do not exist in the public databases, such as children, may be excluded from the data user. However, the process favors false negative correlation over false positive.

A practitioner skilled in the art would recognize the many permutations of the basic concept of the present invention, that is, the use of a trusted third party with a data provider and a data user to depersonalize data as the data passes from provider to user. The embodiments described above are exemplary in nature, and do not constitute an exhaustive listing of the various ways this invention may be implemented.

Figure 7 is a block diagram of an exemplary physical implementation of any of the information networks shown in Figures 1 through 6. The exemplary system is linked by a local area or wide area network 716 which may also be connected to a global information network, such as the Internet, by a direct communications interface 718 and by removable media 722. The exemplary system shown in Figure 7 includes six processing systems, 710, 730, 740, 760, 770 and 780. Each of these systems may include any of the communication interfaces shown for processing system 710. Each of the systems 710, 730, 740, 760, 770 and 780 has an associated database 712, 732, 742, 762, 772 and 782. The databases maintained by the data provider, data user and TTP may reside on any commercially available host computer, as currently known in the art.

The exemplary processing system 710 includes a host computer 714 and a network interface 716 by which the host computer 714 may communicate with other data processing systems via a local area network, a wide area network or a global information network. As shown in Figure 7, the host computer 714 communicates with the processing systems 740 and 730 via a local area network (LAN) 717. Computer 714 also uses the LAN 717 to communicate with a global information network server 750 and, through the server 750 and global information network 752, to remote users 760 and 780. In addition to the network interface, the host computer 714 of the data processing system 710 includes a communications interface 718, for example, a modem, through which the processing system 710 may communicate with the remote user 770. The processing system 710 also includes an input/output (I/O) processor 720 which is coupled to a removable media device 722, for example a diskette drive, through which the host computer can communicate with any other computer system that does not have a direct or indirect data communication path with the host computer 714.

Each host computer may contain one or more processors (not shown), memory (not shown), input and output devices (not shown), and access to mass storage (not shown). Each processing system may be a single system or a network of computers, as currently known in the art. The data providers, TTP and data users may exchange data over computer network such as LAN 717 or by physically transferring data on removable media 722 from location to location. The system may also be implemented across a global information network such as the Internet. The host computer and the global information network may also communicate with a plurality of remote users.

The term "database" may be broadly interpreted to mean any database using records and fields, or their equivalent. The method is not limited by the high-level language used to code the data or the language used to code the programs which implement the required data processing. It is contemplated that the subject invention may be practiced in computer software executed by the data provider(s) 112, trusted third party 116 and data user 118. This computer software may be implemented on a carrier, such as a diskette, CD-ROM, DVD-ROM or radio frequency or audio frequency carrier wave.

Figures 8 and 9 are flow-chart diagrams which illustrate exemplary embodiments of the invention. Figure 8 illustrates a process such as that shown in Figure 6 and Figure 9 shows a process such as that shown in Figures 3, 4 or 5.

In Figure 8, at step 810, the TTP 116 provides the encoding process and encoding database to two retailers, retailer 112a and retailer 112b. The retailers implement the process and database within their company. The databases 115a and 115b provided by the TTP 116 in this exemplary embodiment of the invention are pre-populated with information supplied from the TTP central database 115. The information provided does not include any unique identifiers.

At step 812, each of the retailers 112a and 112b extracts the individual demographic attributes and individual identifiers from each data record it wishes to sent to the data user 118, in this example, a marketing agency. For each record, the information is processed through TTP's supplied encoding process. The encoding process, at step 814 assigns a unique identifier to each record. Next, at step 814, the retailers 112a and 112b create the depersonalized data by replacing the individual demographic attributes and individual identifiers with the single unique identifier provided by the encoding process and send the depersonalized data to the marketing agency 118.

Next, at step 818, the retailers 112a and 112b send, to the TTP 116, the unique identifiers assigned for each record where they encountered a match during the encoding process execution. The TTP 116, at step 820 stores the unique identifier assignment information provided by the retailers 112a and 112b in its central database 115. Also at step 820, the TTP 116 sends the unique identifiers for the retailers 112a and 112b, which link to the same individual, as the correlating information to the marketing agency 118.

At step 822, the marketing agency links the data using the correlating information and performs its marketing study. This study is performed without the ability to identify any individual person. As illustrated by the arrow from block 822 to block 812, the process is iterative. Periodically, the TTP 116 sends updates to the encoding process and database to the retailers 112a and 112b. These updates result from updates / additions to the encoding process central database obtained by TTP 116. After processing these updates, the retailers 112a and 112b send back to the TTP 116 all unique identifiers that were previously assigned by the retailers to the newly supplied information.

It is noted that in this embodiment of the invention, the retailers 112a and 112b never provided any identifiable retail information. The retail data provided by the retailers to the marketing agency had no individual identifiable attributes. Thus, the marketing agency 118 never knew the identity of the actual individuals. Nonetheless, the marketing agency 118 was able to use the power of the retailer's information to enhance marketing study capability.

In the exemplary embodiment of the invention shown in Figure 9, a manufacturer 118 wishes to use the healthcare information of three local healthcare providers to identify the health habits of a specific disease state. Three data providers 112, ProviderA, ProviderB and ProviderC have information which identifies the individual (for example: Member number, social security number, name, etc.). The manufacturer 118, ProviderA, ProviderB and ProviderC contractually authorize a Trusted Third Party (TTP) 116 to encode the healthcare data using the healthcare data encoding process shown in Figure 9.

At step 910 of this process, ProviderA, ProviderB and ProviderC each extracts the individual identifiable information from their internal databases 111 of healthcare records into a file 113. At step 912, ProviderA, ProviderB and ProviderC send the files to TTP 116.

At step 914, the TTP 116 identifies each individual using it's matching process and assigns an Encoding Key to each record. At step 916, the TTP 116 sends the files with the corresponding Encoding Keys back to ProviderA, ProviderB and ProviderC. Next, at step 916, ProviderA, ProviderB and ProviderC replace the individual attributes for each record they wish to send to the manufacturer 118 with the encoding key received from the TTP 116. Also at step 918, ProviderA, ProviderB and ProviderC send the encoded healthcare information files to the manufacturer 118. At step 920, the manufacturer receives the encoded healthcare information files and obtains the correlating data from the TTP 116. Finally, at step 922, the manufacturer 118 links the data from ProviderA, ProviderB and ProviderC and completes its study. It is noted that this study is completed without the manufacturer being able to identify any person.

While the invention has been described in terms of a number of exemplary embodiments, it is contemplated that it may be practiced as described above with variations that are within the scope of the appended claims.

What is Claimed:

1. A method of distributing data records, which include identifying information fields and other data fields, in an information network comprising a data provider, a data user and a trusted third party, wherein the identifying information in each record identifies a person, said method comprising the steps of:

a) separating the identifying information fields from the other data fields for each data record to generate identifying records;

b) transferring a copy of the identifying records to the trusted third party;

c) associating, by the trusted third party, each of the identifying records with a unique identifier, wherein a respectively different unique identifier is assigned to each person identified by one or more of the identifying records; and

d) transferring, by the trusted third party, the unique identifiers to the data provider;

e) associating, by the data provider, the other data fields with the respective unique identifiers to form depersonalized data; and

f) transferring, by each of the data providers, the depersonalized data to the data user.

2. A method according to claim 1 wherein the step of associating the identifying records by the trusted third part includes the step of generating a random identifier that cannot be used to recover any of the identifying information fields as the unique identifier.

3. A method of distributing data records, which include identifying information fields and other data fields, in an information network comprising a plurality of data providers, a data user and a trusted third party, wherein the identifying information in each data record identifies a person, said method comprising the steps of:

a) separating, by each of the data providers, the identifying information fields from the other data fields for each data record to generate identifying records;

b) transferring, by each of the data providers, a copy of the identifying records to the trusted third party;

c) associating, by the trusted third party, each of the identifying records, with a unique identifier, wherein a respectively different unique identifier is assigned to each individual person identified by one or more of the identifying records; and

d) transferring, by the trusted third party, the unique identifiers to the respective data providers from which the identifying records used to generate the unique identifiers were received;

e) associating, by each of the data providers, the other data fields with the respective unique identifiers to form depersonalized data; and

f) transferring, by each of the data providers, the depersonalized data to the data user.

4. A method according to any one of claims 1-3 wherein the step of associating, by the trusted third party, each of the identifying records, with a unique identifier, includes the step of generating a random identifier that cannot be used to recover any of the identifying information fields as the unique identifier, wherein when the identifying information fields provided by more than one of the plurality of data providers corresponds to one person, respectively different unique identifiers are generated for each of the more than one information providers.

5 A method according to any one of claims 1-4 wherein the step of associating, by the trusted third party each of the identifying records, with a unique identifier further includes the steps of:

a) recording, by the trusted third party, a correlation of each person for whom multiple unique identifiers are assigned to form correlating information; and

b) transferring, by the trusted third party, the correlating information to the data user.

6 A method according to any one of claims 1-5 wherein the step of transferring, by the trusted third party, the correlating information to the data user, includes the steps of

a) receiving, from the data user, a request for correlating information for specific ones of the plurality of data providers; and

b) transferring the correlating information for only the specific ones of the plurality of data providers.

7. A method of distributing a plurality of data records, which include identifying information fields and other data fields, in an information network comprising a plurality of data providers, a data user and a trusted third party, wherein the identifying information in each data record identifies a person, said method comprising the steps of:

a) generating, by each of the data providers, a plurality of first unique identifiers from the identifying information fields of the plurality of data records;

b) transferring, by each of the data providers, a copy of the identifying information fields from each of the plurality data records and a respective copy of each of the plurality of unique identifiers, as a respective plurality of identifying records, to the trusted third party;

c) transferring, by each of the data providers, a copy of the other data fields from each of the plurality data records and a respective copy of each of the plurality of first unique identifiers, as a respective plurality of data records, to the data user;

d) associating, by the trusted third party, each of the identifying records, with a second unique identifier, wherein a respectively different second unique identifier is assigned to each individual person identified by one or more of the identifying records; and

e) transferring, by the trusted third party, the first unique identifiers and the second unique identifiers to the data user;

f) associating, by the data user, the other data records provided by the data provider with the unique identifiers provided by the trusted third party.

8. A method of processing and distributing a plurality of data records, wherein each of the plurality of data records contains information used to identify a person, by a trusted third party, said method comprising the steps of:

a) receiving, from a plurality of data providers, a copy of the plurality of identifying records;

b) associating each of the identifying records, with a unique identifier, wherein a respectively different unique identifier is assigned to each individual person identified by one or more of the identifying records;

c) matching records associated with a particular person among the identifying records provided by the plurality of data providers, to generate the second unique identifier which is the same for all identifying records provided by the plurality of data providers, and

d) transferring the unique identifiers to the respective data providers from which the identifying records used to generate the unique identifiers were received.

9. A carrier containing a set of instructions for causing a general purpose computer network comprising a data provider, a data user and a trusted third party, said network accessing a plurality of data records which include identifying information fields and other data fields, wherein the identifying information in each record identifies a person, to perform the following steps:

a) separating the identifying information fields from the other data fields for each data record to generate identifying records;

b) transferring a copy of the identifying records to the trusted third party;

c) associating, by the trusted third party, each of the identifying records with a unique identifier, wherein a respectively different unique identifier is assigned to each person identified by one or more of the identifying records; and

d) transferring, by the trusted third party, the unique identifiers to the data provider;

e) associating, by the data provider, the other data fields with the respective unique identifiers to form depersonalized data; and

f) transferring, by each of the data providers, the depersonalized data to the data user.

10. A carrier according to claim 9 wherein the step of associating the identifying records by the trusted third part includes the step of generating a random identifier that cannot be used to recover any of the identifying information fields as the unique identifier.

11. A carrier containing a set of instructions for causing a network of general purpose computers comprising a comprising a plurality of data providers, a data user and a trusted third party,

accessing a plurality of data records which include identifying information and other fields, wherein the identifying information in each data record identifies a person, said instructions comprising the steps of:

a) separating, by each of the data providers, the identifying information fields from the other data fields for each data record to generate identifying records;

b) transferring, by each of the data providers, a copy of the identifying records to the trusted third party;

c) associating, by the trusted third party, each of the identifying records, with a unique identifier, wherein a respectively different unique identifier is assigned to each individual person identified by one or more of the identifying records; and

d) transferring, by the trusted third party, the unique identifiers to the respective data providers from which the identifying records used to generate the unique identifiers were received;

e) associating, by each of the data providers, the other data fields with the respective unique identifiers to form depersonalized data; and

f) transferring, by each of the data providers, the depersonalized data to the data user

12. A carrier according to claim 11 wherein the step of associating, by the trusted third party, each of the identifying records, with a unique identifier, includes the step of generating a random identifier that cannot be used to recover any of the identifying information fields as the unique identifier, wherein when the identifying information fields provided by more than one of the plurality of data providers corresponds to one person, respectively different unique identifiers are generated for each of the more than one information providers.

13. A carrier containing a set of instructions for causing a network of general purpose computers, said network comprising a plurality of data providers, a data user and a trusted third party, said network accessing a plurality of data records which include identifying information fields and other data fields, wherein the identifying information in each data record identifies a person, to perform a method comprising the steps of:

a) generating, by each of the data providers, a plurality of first unique identifiers from the identifying information fields of the plurality of data records;

b) transferring, by each of the data providers, a copy of the identifying information fields from each of the plurality data records and a respective copy of each of the plurality of unique identifiers, as a respective plurality of identifying records, to the trusted third party;

c) transferring, by each of the data providers, a copy of the other data fields from each of the plurality data records and a respective copy of each of the plurality of first unique identifiers, as a respective plurality of data records, to the data user;

d) associating, by the trusted third party, each of the identifying records, with a second unique identifier, wherein a respectively different second unique identifier is assigned to each individual person identified by one or more of the identifying records; and

e) transferring, by the trusted third party, the first unique identifiers and the second unique identifiers to the data user;

f) associating, by the data user, the other data records provided by the data provider with the unique identifiers provided by the trusted third party.

14. The carrier of claim 13 further comprising instructions to perform the steps of matching records associated with a particular person among the identifying records provided by the plurality of data providers, to generate the second unique identifier which is the same for all identifying records provided by the plurality of data providers, wherein the matching is performed by the trusted third party.

15. A carrier containing a set of instructions for causing a general purpose computer accessing a plurality of data records, wherein each of the plurality of data records contains information used to identify a person, by a trusted third party, to perform the steps of:

a) receiving a plurality of identifying records from a first data provider;

b) associating each of the plurality of identifying records with a unique identifier, wherein a respectively different unique identifier is assigned to each person identified by one or more of the plurality of identifying records; and

c) transferring the unique identifiers to the data provider.

16. A carrier according to claim 15 wherein the step of associating the identifying records includes the step of generating a random identifier that cannot be used to recover any of a plurality of identifying information fields as the unique identifier.

17. A carrier containing a set of instruction for causing a general purpose computer accessing a plurality of data records wherein each of the plurality of data records contains information used to identify a person by a trusted third party, to perform the steps of:

a) receiving, from a plurality of data providers, a copy of the plurality of identifying records;

b) associating each of the identifying records, with a unique identifier, wherein a respectively different unique identifier is assigned to each individual person identified by one or more of the identifying records;

c) matching records associated with a particular person among the identifying records provided by the plurality of data providers, to generate the second unique identifier which is the same for all identifying records provided by the plurality of data providers, and

d) transferring the unique identifiers to the respective data providers from which the identifying records used to generate the unique identifiers were received.

18. A carrier according to claim 17 wherein the step of associating, by the trusted third party, each of the identifying records, with a unique identifier, includes the step of generating a random identifier that cannot be used to recover any of the identifying information fields as the unique identifier, wherein when the identifying information fields provided by more than one of the plurality of data providers corresponds to one person, respectively different unique identifiers are generated for each of the more than one information providers.

Figure 1

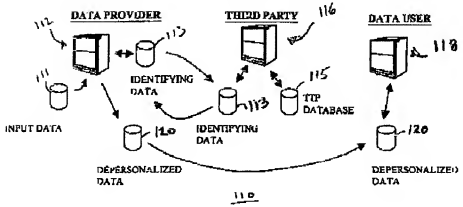


Figure 2

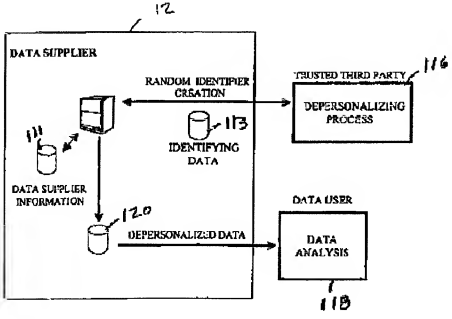


Figure 3

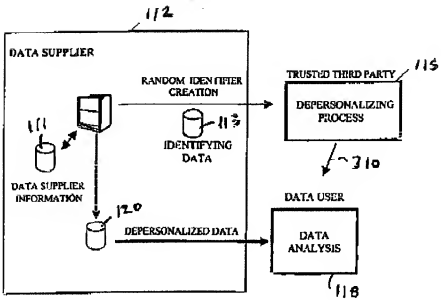


Figure 4

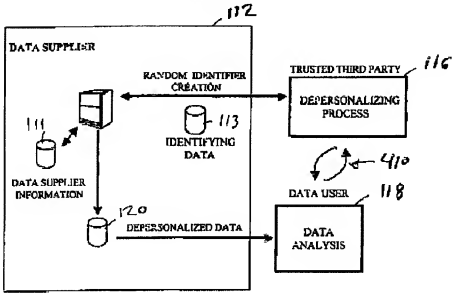


Figure 5

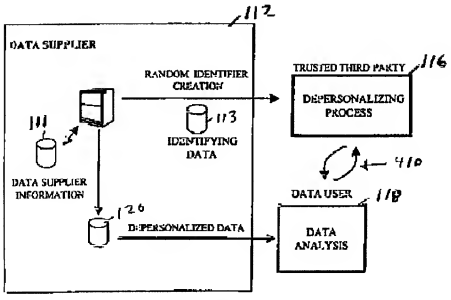


Figure 6

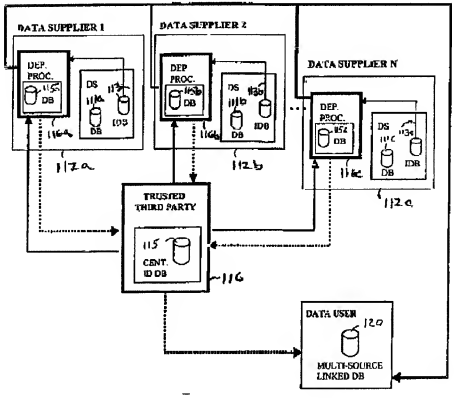


Figure 7

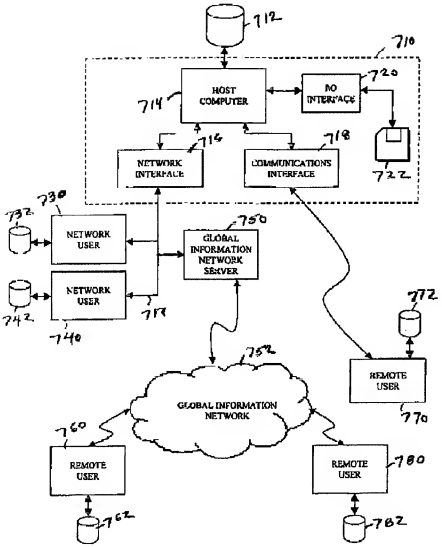

712
710
HOST COMPUTER
714
I/O INTERFACE
720
716
718
NETWORK INTERFACE
COMMUNICATIONS INTERFACE
722
730
732
NETWORK USER
750
GLOBAL INFORMATION NETWORK SERVER
NETWORK USER
742
740
718
772
REMOTE USER
770
752
GLOBAL INFORMATION NETWORK
760
REMOTE USER
780
762
782

Figure 8

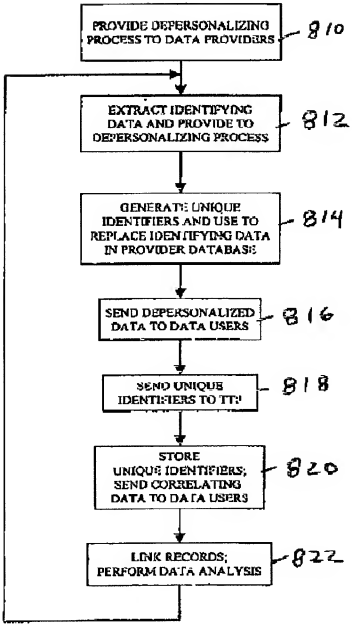
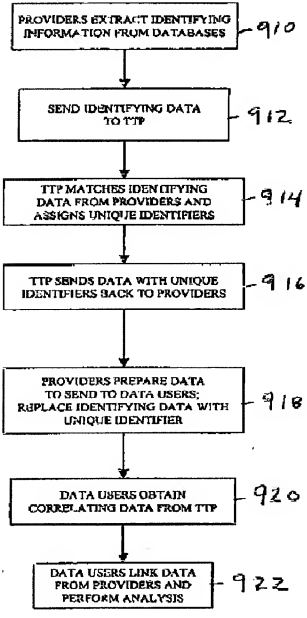

PROVIDE DEPERSONALIZING PROCESS TO DATA PROVIDERS
810
EXTRACT IDENTIFYING DATA AND PROVIDE TO DEPERSONALIZING PROCESS
812
GENERATE UNIQUE IDENTIFIERS AND USE TO REPLACE IDENTIFYING DATA IN PROVIDER DATABASE
814
SEND DEPERSONALIZED DATA TO DATA USERS
816
SEND UNIQUE IDENTIFIERS TO TTP
818
STORE UNIQUE IDENTIFIERS; SEND CORRELATING DATA TO DATA USERS
820
LINK RECORDS; PERFORM DATA ANALYSIS
822

Figure 9


PROVIDERS EXTRACT IDENTIFYING INFORMATION FROM DATABASES
910
SEND IDENTIFYING DATA TO TTP
912
TTP MATCHES IDENTIFYING DATA FROM PROVIDERS AND ASSIGNS UNIQUE IDENTIFIERS
914
TTP SENDS DATA WITH UNIQUE IDENTIFIERS BACK TO PROVIDERS
916
PROVIDERS PREPARE DATA TO SEND TO DATA USERS; REPLACE IDENTIFYING DATA WITH UNIQUE IDENTIFIER
918
DATA USERS OBTAIN CORRELATING DATA FROM TTP
920
DATA USERS LINK DATA FROM PROVIDERS AND PERFORM ANALYSIS
922

Abstract

A computer implemented method allows an owner or provider of data that contains personal identifiers (data provider) to distribute that data to a data user in a depersonalized form, i.e., without revealing the identity of the individuals associated with the data. The data provider first separates the personal information from the other data to create two data sets. The personal identifying information is then provided to a Trusted Third Party (TTP). The TTP associates a unique identifier with the identifying information. This unique identifier replaces any data in the database that can be used to identify an individual, such as name, address or social security number. The TTP may also collect and store the personal identifying information so that it can process identifying information that it acquires in the future to determine if the unique identifiers generated by the data provider or by the TTP refer to the same individual. The data provider associates its own unique identifier or the identifier provided by the TTP with the other data to create depersonalized data that may be sent to a data user for analysis. In this manner, different records from one or more data providers that refer to a single individual can be matched by the data user, and the data provider is assured that no personal identifying information is distributed that would link an individual to a particular data record. The TTP transmits information that correlates unique identifiers from multiple data providers to a data user. Each data provider transmits the depersonalized data, including the unique identifiers to the data user. The data user correlates the information from the different data providers before analyzing the data.